20th INTERNATIONAL CONFERENCE

ON SCIENTOMETRICS & INFORMETRICS

ISSI 2025

23–27 June 2025

Yerevan, Armenia

PROCEEDINGS

Editors

Shushanik Sargsyan, Wolfgang Glänzel, Giovanni Abramo

UDC 001.8(082)

Sponsors

Diamond sponsor – Higher Education and Science Committee, RA MoESCS

Gold sponsor - Journal of Data and Information Science

Silver sponsor – Clarivate

Bronze sponsor – Quantitative Science Studies

Bronze sponsor – Springer

Partners

Institute for Informatics and Automation Problems of NAS RA

Yerevan State University

Center for Scientific Information Analysis and Monitoring

ISBN 978-9939-1-2086-7

ISSN 2175-1935

June 2025 Printed in Armenia

Organizing Committee

Local Members

Shushanik Sargsyan, Armenia, *Chair*

Edita Gzoyan, Armenia

Aram Mirzoyan, Armenia

Gevorg Kesoyan, Armenia

Simon Hunanyan, Armenia

Yeranuhi Manukyan, Armenia

International Members

Jacqueline Leta, Brazil

Vivek Kumar Singh, India

Program Committee Chairs

Sargis Hayotsyan, Armenia

Hovhannes Hovhannisyan, Armenia

Hrachya Astsatryan, Armenia

Giovanni Abramo, Italy

Wolfgang Glanzel, Belgium

Doctoral Forum Committee

Iana Atanassova, France

Andrea Scharnhorst, Netherlands

Gunnar Sivertsen, Norway

Workshops & Tutorials Committee

Cinzia Daraio, Italy

Alesia Zuccala, Denmark

Poster Session Committee

Jacqueline Leta, Brazil

Pei-Shan Chi, Taiwan, Belgium

Eugene Garfield Award Committee

Guillaume Cabanac, France

Nees Jan van Eck, Netherlands

Mike Thelwall, United Kingdom

Lin Zhang, China, Belgium

Student Travel Award Committee

Dag W. Aksnes, Norway

Rodrico Costas, Netherlands

Juan Gorraiz, Austria

Best Paper Award Committee

Nicolas Robinson Garcia, Spain

Vivek Singh, India

Cassidy Sugimoto, USA

Scientific Committee

Dag W. Aksnes, Norway

Iana Atanassova, France

Alberto Baccini, Italy

Rafayel Barkhudaryan, Armenia

Aparna Basu, India

Marc Bertin, France

Sujit Bhattacharya, India

Kevin Boyack, USA

Guillaume Cabanac, France

Zaida Chinchilla-Rodríguez, Spain

Ciriaco Andrea D'Angelo, Italy

Cinzia Daraio, Italy

Gemma Derrick, United Kingdom

Sergio Luiz Monteiro Salles Filho, Brazil

Wolfgang Glänzel, Belgium

Claudia Gonzalez-Brambila, Mexico

Maria Cláudia Cabrini Gracio, Brazil

Edita Gzoyan, Armenia

Robin Haunschild, Germany

Hamid R. Jamali, Australia

Vincent Larivière, Canada

Jacqueline Leta, Brazil

Domenico Augusto Maisano, Italy

Philipp Mayr, Germany

Rogério Mugnaini, Brazil

Bosire Onyancha, South Africa

Gangan Prathap, India

Emanuela Reale, Italy

Ronald Rousseau, Belgium

Shushanik Sargsyan, Armenia

Vivek Kumar Singh, India

Gunnar Sivertsen, Norway

Cassidy Sugimoto, USA

Mike Thelwall, United Kingdom

Preface

It is our great pleasure to present the *Proceedings of the 20th International Conference on Scientometrics & Informetrics (ISSI2025)* of the International Society for Scientometrics and Informetrics, held from June 23 to 27, 2025, in Yerevan, Armenia. This edition of the ISSI conference marks four decades of global exchange and collaboration in the field of scientometrics and informetrics — a field that continues to grow in relevance as science itself evolves in complexity, scope, and global impact.

Hosted for the first time in the South Caucasus, ISSI2025 brought together over 230 participants from more than 38 countries, making it one of the most geographically and thematically inclusive gatherings in the history of the ISSI community. The conference's theme — "Shaping the Future: New Horizons in the Science of Science" — inspired reflection on our field's legacy while encouraging the exploration of bold new directions for scientometric research.

The conference opened with warm welcomes from local and international leaders, including representatives of Armenia's academic and governmental institutions and the President of ISSI.

Two keynote addresses, delivered by **Mike Thelwall** and **Gunnar Sivertsen**, focused on some of the field's most burning questions — including the use of Large Language Models in evaluative contexts and the core values guiding our work in research assessment and policy-relevant application.

During five days, ISSI2025 featured:

- **More than 30 parallel sessions,** showcasing cutting-edge work in areas such as advanced informetric models, science policy and research evaluation, artificial intelligence and scientometrics, open science, micro- and macro-level analysis, technology and innovation studies, and gender, collaboration, and mobility in science.
- **The Doctoral Forum**, providing early-stage researchers an opportunity to present and discuss their work with peers and senior experts in the field.
- **Workshops and Tutorials**, including:
    - The Joint Workshop of the 5th AI + Informetrics (AII) and the 6th Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE) — AII-EEKE 2025
    - The tutorial "Exploring the OpenAIRE Graph on Google Big Query", offering hands-on insights into open scholarly data
- **Two special tracks:**
    - *FRAME* (*Framework for the Responsible Use of Assessments and Metrics in Evaluation*), dedicated to developing fair, inclusive, and context-sensitive approaches to research evaluation

    - *Open Research Information (ORI)*, focused on infrastructures and practices in sharing scientific metadata
- A **poster session**, featuring a wide range of emerging work and interdisciplinary projects
- The presentation of the prestigious **Derek de Solla Price Award** by the international journal *Scientometrics*
- **Student Travel Awards**, supporting young researchers from around the world
- An award ceremony and closing events, celebrating contributions from across the global scientometric community, including:
    - The **Eugene Garfield Doctoral Dissertation Scholarship**, awarded to an exceptional doctoral student for outstanding research in the field
    - **Best Paper Award**, recognizing the most impactful and innovative research presented at the conference
- The **ISSI General Assembly**, where future plans and institutional developments were discussed

This volume publishes the peer-reviewed papers (*full papers, research-in-progress, and poster papers*) presented during the conference. It reflects the thematic richness and methodological diversity of our community, and highlights the increasing interrelation between scientometric methods and broader societal challenges — including sustainability, policy innovation, and the responsible use of AI.

We would like to express our deepest thanks to the International Society for Scientometrics and Informetrics (ISSI) for their trust and support, as well as to our academic partners, sponsors, and institutional collaborators in Armenia. Special appreciation goes to the reviewers, session chairs, keynote speakers, and the tireless members of the organizing and scientific committees.

Most importantly, we thank all authors and participants for their contributions to this vibrant intellectual exchange. May these proceedings serve as a valuable resource for ongoing research, and as inspiration for the continued development of scientometrics as a field committed to rigor, openness, and global inclusivity.

Shushanik Sargsyan, Wolfgang Glänzel, Giovanni Abramo

Index of proceedings papers

Framework for the Responsible use of Assessments and Metrics in Evaluation (Frame)

Open Research Information (ORI)

Full papers

Special Track: Framework for the Responsible use of Assessments and Metrics in Evaluation (FRAME)

https://doi.org/10.51408/issi2025_159

Advancing Responsible Bibliometric Practices in Research Assessment

An Introduction to the ISSI 2025 Special Track "A framework for the responsible use of bibliometrics in research evaluation" (FRAME)

Cinzia Daraio[1], Wolfgang Glänzel[2], Juan Gorraiz[3]

[1]*daraio@diag.uniroma1.it*
DIAG, Sapienza University of Rome (Italy)

[2]*wolfgang.glanzel@kuleuven.be*
ECOOM, KU Leuven (Belgium)

[3]*juan.gorraiz@univie.ac.at*
Dept Bibliometrics & Publication Strategies, University of Vienna (Austria)

The role of bibliometric indicators in research evaluation has undergone substantial evolution over the past decades, becoming integral to institutional assessments, funding decisions, and science policy at large. While their widespread adoption has enabled new forms of analysis and benchmarking, it has also sparked ongoing debates around their transparency, ethical integrity, and contextual relevance. A central concern is the over-reliance on narrow performance metrics, such as publication counts or citation-based rankings, often applied uniformly and without sufficient consideration of disciplinary norms, research diversity, or the broader societal value of scientific work.

Recent international initiatives—most notably the *Agreement on Reforming Research Assessment* (CoARA, 2022)—have brought renewed attention to the need for holistic, inclusive, and responsible evaluation frameworks. These efforts underscore the importance of balancing quantitative indicators with qualitative judgment, and of ensuring that assessment systems reinforce, rather than distort, the values of academic integrity, transparency, and societal engagement.

In this context, the special track "*A Framework for the Responsible Use of Bibliometrics in Research Evaluation*" (FRAME), hosted at ISSI 2025, aims to foster critical reflection and advance practical guidance for the responsible integration of bibliometrics into research assessment. Its overarching goal is to articulate concrete criteria, protocols, and governance models that ensure metrics are used ethically, appropriately, and effectively across diverse evaluation settings.

This track is structured around five core objectives:

1. To promote a shared understanding of what constitutes responsible metrics, clarifying their scope of application and potential limitations.
2. To co-design evaluation systems that align with principles of academic rigor, respond to the needs of diverse stakeholders, and account for both scholarly and societal impact.

3. To develop standardized protocols and ethical guidelines to improve the transparency, reproducibility, and inclusivity of metric-based assessments.
4. To anticipate and address challenges posed by emerging technologies, particularly the increasing use of AI tools in scientific publishing, reporting, and peer review.
5. To foster collaboration among key actors in the scientometric community, including researchers, practitioners, funders, and policy institutions.

To support these goals, the FRAME track invited interdisciplinary contributions that explore the theoretical foundations, practical applications, and policy implications of responsible bibliometric use.

The selected papers include:

- Conceptual frameworks and empirical models that support context-sensitive indicator use (Daraio, Glänzel, and Gorraiz, 2025; Xenou et al., 2025).
- Studies evaluating the interplay between academic and non-academic impact in assessment models (Haunschild and Bornmann, 2025).
- Analyses of how bibliometric indicators influence research behavior, institutional strategies, and policy formulation (Engels., Houben and Spooren, 2025; Rousseau, 2025; Sivertsen, Zhang, and Rushforth, 2025).
- Reforming research assessment in Social Science and Humanities (SSH, Bonaccorsi, 2025).
- Critical assessments of AI-assisted tools in scholarly communication and their implications for research evaluation systems (Thelwall 2025).
- Ethical, legal, and social considerations surrounding indicator selection, data governance, and accountability (Huang et al., 2025; Kosmulski, 2025; Tejada-Gómez and Ayure-Urrego, 2025; Vaccari and Daraio, 2025).

By convening diverse perspectives and encouraging methodological innovation, this track contributes to the broader agenda of transforming research assessment systems in ways that are fair, credible, and future-oriented. It supports the development of actionable tools and shared principles that enable responsible metric use while embracing the complexity and diversity of contemporary scientific practice.

List of contributions to the Special Track

Bonaccorsi A. (2025), "The new alliance. Bringing together bibliometric and library science for a responsible assessment of research in SSH", in this Special Track.

Daraio C., Glänzel W., Gorraiz J. (2025), A Responsible Framework for an Appropriate Bibliometric-Based Research Assessment, in this Special Track.

Engels T. C. E., Houben B., Spooren P. (20205), Responsible research assessment of teams: reflections and perspectives after two evaluation cycles at the University of Antwerp, Belgium, in this Special Track.

Haunschild R., and Bornmann L. (2025), Mapping national research that targets sustainable development goals: The responsible visualization of OpenAlex data for societal impact measurements of research, in this Special Track.

Huang Y., Liu W., Fu H., Ma J., Zhang G., Bu Y., Min C., Wu Z. (2025), Toward Responsible Scientometrics: Normative Data Practices for Research Evaluation, in this Special Track.

Kosmulski M. (2025), Can scientific papers be unretracted? in this Special Track.
Rousseau R. (2025), Trueblood et al.'s Ideas on Research Evaluation and Implications for Reforming Research Assessment, in this Special Track.
Sivertsen G., Zhang L., Rushforth A. (2025), "Responsible metrics for the assessment of research organizations", in this Special Track.
Tejada-Gómez M. A., Ayure-Urrego M. (2025), Ethical and responsible model for the National Science, Technology and Innovation System in Colombia, in this Special Track.
Thelwall M. (2025), Responsible Uses of Large Language Models for Research Evaluation, in this Special Track.
Vaccari A., Daraio, C. (2025), Does Evaluating Research Still Need Virtues in the Age of Chat GPT?, in this Special Track.
Xenou Z., Malanguarneral G., Provost L., Manola N. (2025), Towards a Responsible Research Assessment Transition: A Novel Framework for Researcher Profiles, in this Special Track.

https://doi.org/10.51408/issi2025_160

A Responsible Framework for an Appropriate Bibliometric-Based Research Assessment

Cinzia Daraio[1], Wolfgang Glänzel[2], Juan Gorraiz[3]

[1]*daraio@diag.uniroma1.it*
DIAG, Sapienza University of Rome (Italy)

[2]*wolfgang.glanzel@kuleuven.be*
ECOOM, KU Leuven (Belgium)

[3]*juan.gorraiz@univie.ac.at*
Dept Bibliometrics & Publication Strategies, University of Vienna (Austria)

Abstract

The growing reliance on bibliometric indicators in research evaluation has generated increasing criticism, both from the academic community and recent European initiatives advocating more holistic, peer review–centered approaches. This paper addresses the urgent need for responsible and contextualised use of such metrics. Rather than rejecting bibliometrics completely, we propose a conceptual framework that supports the appropriate application of bibliometric indicators, tailored to the goals, disciplinary contexts, and levels of analysis involved. This framework promotes a balanced approach, valuing transparency, interpretive care, and ethical use of quantitative indicators within broader evaluation systems. The paper, interpreting and substantiating CoARA (2022)'s claims, emphasises the integration of metrics with qualitative assessments to ensure academic integrity and societal relevance. It calls for shared protocols, cross-sector collaboration, and recognition of disciplinary diversity to ensure indicators *inform* rather than disappear from research assessment or dominate research assessment.

Introduction

In recent years, the application of quantitative approaches – particularly bibliometric indicators – in research assessment has come under intense scrutiny. Much of this criticism stems from concerns over the unintended consequences of these tools when used improperly. However, reform initiatives often lack conceptual clarity: they seldom define what exactly is being evaluated, at which level of aggregation, and with what granularity. This ambiguity leaves open whether "research" refers to a holistic academic process or merely to measurable outputs. Complicating matters further, much of the critique favoring peer review over metrics is based on issues observed at the individual researcher level – problems already acknowledged within the bibliometric community itself (Wouters et al., 2013).

This skepticism towards indicators has spurred a wave of manifestos and declarations—such as DORA and the Leiden Manifesto – advocating for more responsible and meaningful approaches to research assessment (Wilsdon et al. 2015; Biagioli and Lippman, 2020; Curry et al., 2020).

At the European policy level, calls for change have intensified. The European Commission's 2021 scoping report advocates for a re-evaluation of current systems and was foundational for the CoARA agreement in July 2022 (European

Commission, 2021; CoARA, 2022). While these initiatives mark significant progress, they do not offer concrete operational tools or criteria for responsible indicator use (see also Daraio and Maletta, 2025).

In response, this paper argues that bibliometric indicators should not be dismissed altogether. Rather, their "inappropriate" use – such as applying them in contexts for which they were never intended – should be the real target of reform (Glänzel, 2006). Bibliometric indicators are analytical tools developed through rigorous scientific methods within the fields of scientometrics and information science. Hence, discrediting them broadly is both unjustified and counterproductive.

What is required is a structured framework to determine whether the use of a specific indicator is fit-for-purpose in each evaluation context. The goal is not to oppose quantitative methods with qualitative ones, but to develop criteria that guide appropriate use, acknowledging that even peer review has limitations.

Thus, the paper proposes a multidimensional framework that outlines how indicators should be selected and applied responsibly in varying evaluation contexts. It concludes by identifying critical questions and limitations, while affirming the value of indicators – when used with expertise and care – in contemporary research evaluation.

Key Framework Dimensions for Evaluative Bibliometrics

In an influential contribution, Henk Moed (2017) introduced a visionary model of "evaluative informetrics," emphasizing how to practically apply bibliometric methods in research assessment. He later refined this framework, outlining the following four central questions essential to shaping evaluation studies.

1. What is the unit of assessment (e.g., individual, institution, country)?
2. What aspect of the research process is under consideration (e.g., scholarly impact, social benefit, interdisciplinarity, collaboration)?
3. What are the goals of the evaluation (e.g., resource allocation, performance improvement, strategic redirection)?
4. What are the characteristics of the assessed entities, including developmental stage or systemic relevance (Moed, 2020, p. 4)?

Table 1. The six dimensions of our Research Evaluation Framework.

#	Dimension	Definition	Warnings (or Pitfalls)
1	*Aggregation Level*	The scale at which evaluation is conducted: individual, group, institution, region, or nation.	Metrics must match the level: those valid at one level may mislead at other. Peer review suitability decreases with higher aggregation.
2	*Unit of Assessment*	The specific entity or profile being evaluated (e.g., individual researcher, lab, department).	Influenced by the context and nature of research; discipline and sector-specific needs matter.

3	*Purpose of Assessment*	The goal of the evaluation, such as funding, improvement, promotion, or benchmarking.	Drives methodology, timeline, baseline, and criteria. Different objectives call for different evaluation strategies.
4	*Context of Assessment*	The broader environment, conditions, and institutional or national environment in which research takes place.	Evaluations must be sensitive to systemic, geographical, or disciplinary contexts to avoid bias or misinterpretation.
5	*Elements of Research Process*	The stages and outputs of research, including input, process, output, and impact (academic and non-academic).	Must consider diverse impacts (e.g., social, economic, cultural) beyond scholarly output.
6	*Stakeholder Engagement*	Inclusion of those affected by or involved in research and evaluation: funders, institutions, public, etc.	Helps assess broader impact and legitimacy of evaluation; considers intended and unintended consequences.

Building on this foundation, we propose an expanded multidimensional framework by introducing two additional dimensions (see Daraio et al., 2024). Table 1 gives an overview of the proposed dimensions.

Criteria for Building and Using Research Evaluation Metrics Appropriately

The use of bibliometric and other quantitative indicators in research evaluation has grown increasingly complex. As shown in the Multidimensional Research Assessment Matrix (AUBR, 2010) and expanded by Moed (2017), there exists a broad array of indicators and methods intended to assess both scholarly and non-academic research impacts. While Moed (2017) offers concrete recommendations and evaluations of specific metrics, the AUBR matrix provides a more general overview of methods and their potential applications.

However, simply selecting from existing indicators is not enough. Even scientifically sound and well-designed metrics can lead to harmful conclusions if applied out of context. Therefore, the focus should not only be on building reliable metrics, but also on ensuring their appropriate application—tailored to the specific goals, level of aggregation, and nature of the research being evaluated.

To combine quantitative and qualitative methods meaningfully, diverse data types must be harmonized. Daraio and Glänzel (2016) proposed a standardized data integration model to support this process.

For bibliometric indicators to be meaningful and robust, they must meet several foundational conditions: i) data quality is essential; ii) metrics must ensure comparability (*commensurability*) and iii) results should be replicable over time (*validatability*). Bookstein (1997) further warns that measurement efforts are often undermined by randomness, ambiguity, and conceptual fuzziness. These challenges affect both metric design and interpretation.

To be considered fit for research assessment, indicators must meet a set of core criteria: they must be valid, meaningful, reliable, robust, and, where possible, normalisable and standardisable. This ensures that indicators are suitable for comparative evaluations and benchmarking.
Even after rigorous design, indicators must be applied within a conceptual framework that accounts for:

- The unit of analysis (e.g., researcher, institution),
- Disciplinary differences,
- Data infrastructure and publication behaviour.

Importantly, metrics must be selected based on their "fitness for purpose" – their ability to align with the specific assessment goals. Users should be aware of the margins of error they are willing to tolerate and interpret results in light of possible limitations or methodological flaws.
While both ex-ante and ex-post assessments are valuable, they require different types of data and interpretation. Therefore, a thoughtful balance between qualitative and quantitative approaches is essential. Qualitative aspects – like recognition, diversity, and societal engagement – must not be overlooked.
Finally, caution is advised when using composite indicators, which often suffer from non-transparency, arbitrary weighting, and component interdependence. Their tendency to compress multidimensional realities into a single value may obscure more than it reveals.
Table 2 offers a concise yet comprehensive overview of key dimensions that must be considered to ensure responsible, meaningful, and context-sensitive use of bibliometric indicators. It emphasizes that indicators should not be applied in isolation, but rather aligned with the purpose, unit of assessment, disciplinary norms, and stakeholder perspectives. By explicitly addressing methodological, interpretive, and ethical concerns – such as data quality, transparency, and fitness for purpose – the table supports evaluators in navigating complex assessment environments. It could be useful as a *practical checklist* or *diagnostic tool* to guide the informed and balanced application of metrics within broader evaluation frameworks.

Table 2. Criteria for the appropriate use of indicators in research evaluation.

Criteria	Key Elements/ Insights	Sources
1. Foundational Frameworks	- AUBR Matrix (2010) outlines multi-dimensional methods for assessing research performance. - Moed's evaluative informetrics (2017) provides practical applications, distinguishing academic and non-academic impacts. - Extends to alternative metrics for broader impacts.	AUBR (2010); Moed (2017)

2. *Appropriate Use*	- Indicators must be contextually appropriate – not all are fit for all settings. - Even valid metrics can mislead or harm when used improperly. - Importance of selecting metrics aligned with evaluation goals, level of aggregation, and disciplinary context.	General argument from paper; Glänzel (2006); Gorraiz et al. (2020)
3. *Data Integration*	- Combining qualitative and quantitative approaches requires harmonizing different types of data. - Standardized integration model proposed by Daraio & Glänzel (2016) to support coherent use of data in multi-purpose assessments.	Daraio & Glänzel (2016)
4. *Basic Data Requirements*	- *Quality***:** Data must be accurate, verified, and trustworthy. - *Commensurability*: Enables comparisons across cases, institutions, or disciplines. - *Validatability*: Results must be reproducible under identical data collection conditions.	Daraio & Glänzel (2016); Bookstein (1997)
5. *Measurement Pitfalls*	- *Randomness*: Unpredictable variability in measurement. - *Fuzziness*: Lack of clear definition or conceptual sharpness. - *Ambiguity*: Interpretational uncertainty. These issues affect both metric design and interpretive clarity.	Bookstein (1997)
6. *Indicator Criteria*	Indicators should be: - *Valid* – Measures what it claims to measure. - *Meaningful* – Yields interpretable, relevant insights. - *Reliable* – Statistically stable and reproducible. - *Robust* – Insensitive to minor changes in the system. - *Normalisable* – Adaptable to different scales. - *Standardisable* – Comparable and replicable across contexts. - *Quality-based* – Depends on high-quality data sources.	Moed (2017); Bookstein (1997); Daraio & Glänzel (2016); Gorraiz et al. (2016)
7. *Application Considerations*	Indicators must be aligned with - The *unit of assessment* (individual, institution, etc.) - The discipline's characteristics (e.g., citation practices) The *purpose of the evaluation* (e.g., funding, promotion) - Available infrastructure, data, and evaluation goals.	Moed (2017); EU Scoping Report (2021)

Table 2 (*contd*.). Criteria for the appropriate use of indicators in research evaluation.

Criteria	Key Elements/ Insights	Sources
8. Composite Indicators Warning	Should be used with caution**:** - Tend to obscure complexity. - May rely on *arbitrary weightings* and *inconsistent metrics*. - Risk loss of transparency**,** misinterpretation, and over-simplification. Interdependence of components may introduce systemic bias or noise.	General critique from paper; Moed (2017)
9. Balancing Methods	Responsible evaluation requires combining metrics with: - Peer review and expert input - Narratives and case-based evidence - Qualitative factors like diversity, recognition, and societal impact. Ensures fairness, inclusivity, and relevance across varied contexts.	Moed (2007); Best practice in research evaluation literature
10. Responsible use of indicators in research assessment	- Indicators must be applied with awareness of their limitations, context-dependence, and potential unintended consequences. - Requires critical reflection on indicator selection, data quality, purpose alignment, and fairness. - Must avoid mechanistic or symbolic use of metrics (e.g., compliance without reform). - Emphasizes transparency, reproducibility, stakeholder engagement, and ethical responsibility in interpretation and application. - Encourages use of indicators as *decision-support* tools, not decision-makers.	CoARA (2022); EU (2021); Moed (2017); Curry et al. (2020); General principles from the paper

An illustration of our framework

Figure 1 illustrates our framework that can be represented by an *optical prism*: The Prism of Research Evaluation. Just as a prism refracts white light into a spectrum of colours, the prism in this figure refracts the "light" of research performance through a structured and multi-dimensional evaluative lens. This figure signals a fundamental principle in responsible assessment: research quality is not a single colour or metric, but a multifaceted, context-sensitive construct. The three basic dimensions of our framework, the basis of our prism in Figure 1, from which to begin are: the unit to be evaluated (*whom we are assessing*), the research process to be evaluated considering its boundaries (*what we are assessing*), and the main goal of the assessment (*why we are doing the assessment*). We then have two important dimensions that allow us to specify *where*, *when*, and most importantly, *how* the assessment is carried out. They are the level of aggregation and the context of the evaluation, which constitute the two sides of our framework. Finally, we have the

dimension that completes our framework represented by all *stakeholders* interested in the evaluation and its impacts and effects (consequences). Our framework aims to apply some kind of spectral decomposition of the complex assessment task represented by light entering the prism. If it works correctly, the prism should provide a proper evaluation spectrum for the unit under assessment.

The refracted rainbow from the prism signals the diverse outcomes of evaluation when it is performed responsibly. No single metric or ranking can capture this plurality. Instead, we must strive to view research through multiple lenses, acknowledging that different purposes and contexts will yield different "colours" of insight.

This model embodies several key elements of responsible evaluation:

- *No one-size-fits-all*: Good assessment requires contextual fit between indicators and purpose.
- *Critical reflection*: Encourages evaluators to think through the boundaries and assumptions that structure assessment.
- *Participatory governance*: Promotes involvement of all stakeholders in defining meaningful metrics and methods.
- *Transparency*: Reveals how decisions are derived and reduces the black-boxing of evaluative procedures.
- *Indicator pluralism*: Supports a multidimensional approach to research assessment.

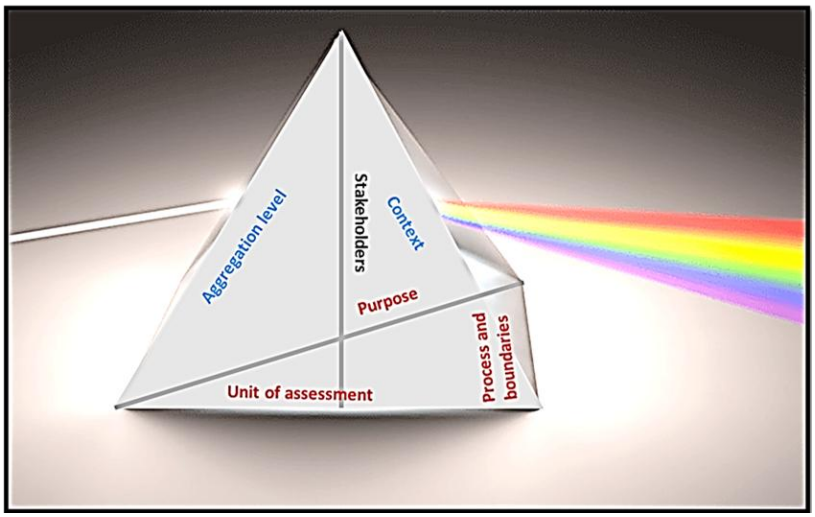


Figure 1. The Prism Model of Responsible Research Assessment.

Conclusions – Responsible and Contextual Use of Indicators in Research Evaluation

To ensure that bibliometric indicators are applied responsibly, a robust framework is essential – one that guides users in selecting the most appropriate metrics based on the specific goals, context, and evaluation problem at hand. The framework proposed in this paper aims to assist evaluators in choosing indicators that are fit for purpose

and in determining acceptable levels of uncertainty or error depending on the evaluation context. It also encourages the development of checklists to match available indicators to key assessment dimensions, thereby promoting structured and transparent decision-making (Robinson et al., 2024).

However, using the right indicators is not enough; they must be interpreted critically and carefully, with full awareness of the limitations imposed by methodological weaknesses, data quality, and parameter selection.

The shift from "publications and citations" to a broader spectrum of research contributions raises important concerns. What alternative outputs should be included in evaluation? How can we ensure these do not replicate the very problems that traditional citation-based metrics introduced – such as encouraging quantity over quality? For example, if researchers are evaluated based on uploaded outputs rather than impactful contributions, similar forms of metric manipulation could emerge. One proposal to counteract this issue, limiting the number of outputs submitted for evaluation, could help restore quality-based incentives. Yet such policies must be designed carefully to avoid unintended effects, such as disadvantaging early-career researchers or disciplines with rapid publication cycles.

Transparency and reproducibility must remain core principles in all evaluation methodologies. These can only be achieved if indicator use is standardized, well-documented, and paired with regular stakeholder interaction, including with researchers, institutions, and the wider community. Such engagement enhances both the meaningfulness and accuracy of the evaluation process and helps in identifying acceptable error thresholds and interpretive caveats.

To meet the complexity of today's research environment, bundles of valid and robust indicators should be selected, not created by arbitrarily combining metrics into opaque composite indicators. The paper cautions against composite indicators, as they often distort multi-dimensional realities, force linearity, and reduce transparency and interpretability. These effects directly conflict with the core principles of responsible metric use.

Recent approaches such as "narrative bibliometrics" (Torres-Salinas et al., 2024) offer a promising alternative. By embedding bibliometric data within contextualized, narrative interpretations, this method can enrich our understanding of impact, especially for less easily quantified outputs. Yet this, too, comes with limitations. The shift from objective metrics to subjective narratives introduces interpretive variability, which may undermine the neutrality typically associated with bibliometrics.

As Moed (2007) highlighted, the most effective evaluations combine "advanced metrics" with "transparent peer review". However, just as metrics require clear criteria for validity and reliability, qualitative evaluations also face challenges. Biases such as arbitrariness and fuzziness, critiqued by Bookstein (1997) in quantitative contexts, can also be present in peer review and narrative assessments.

Lastly, the growing role of Artificial Intelligence (AI) in bibliometrics introduces both opportunities and risks. AI tools can enhance data interpretation, detect meaningful patterns, and automate large-scale analyses. But they also risk reinforcing algorithmic biases, reducing human oversight, or narrowing the

evaluation lens. Any AI-based tools must be deployed with strong ethical guardrails, human interpretability, and accountability.

Rethinking research assessment is a complex but necessary undertaking. Incorporating a diversity of research outputs, improving the appropriateness of metric use, and embedding evaluation in ethical, transparent, and participatory practices are all critical. Achieving this will require not just methodological innovation, but active collaboration among researchers, institutions, funders, and policymakers.

Acknowledgement

This contribution is based on the conference paper by Daraio et al. (2024) presented as part of a special session organised by the authors at the STI 2024 Conference in Berlin. The objective of the present study is to contribute to the continuation of the debate within the framework of a special track at ISSI 2025 in Yerevan.

References

AUBR (2010), *Assessing Europe's University-Based Research – Expert Group on Assessment of University-Based Research*. (2010). Research Policy. European Commission. doi:10.2777/80193

Biagioli M, Lippman A eds. (2020), *Gaming the metrics: Misconduct and manipulation in academic research*, MIT Press.

Bookstein A. (1997), Informetric distributions. III. Ambiguity and randomness. *JASIS*, 48(1), 2–10.

CoARA (2022), *Coalition for Advancing Research Assessment*. (accessible at: https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf)

Curry, S., de Rijcke, S., Hatch, A. et al. (2020), The changing role of funders in responsible research assessment: progress, obstacles and the way ahead. Working Paper. Research on Research Institute (RoRI) https://doi.org/10.6084/m9.figshare.13227914.v1

Daraio, C., Glänzel, W. (2016). Grand challenges in data integration—State of the art and future perspectives: An introduction. Scientometrics, 108(1), 391-400.

Daraio, C., Glänzel, W. (2020). Selected essays of Henk F. Moed. *Evaluative Informetrics: The Art of Metrics-Based Research Assessment: Festschrift in Honour of Henk F. Moed*, 15-67.

Daraio C., Gorraiz J., Glänzel W. (2024), Towards a framework for the appropriate use of bibliometric indicators in research evaluation, STI 2024 Conference, 18-20 September 2024, Berlin (Germany), https://doi.org/10.5281/zenodo.14036242.

Daraio C., Maletta S. (2025), Understanding Responsible Research Assessment: A MacIntyrean Proposal, *Acta Philosophica, International Journal of Philosophy*, 1, 34, 173-188.

EU (2021), "*Towards a reform of the research assessment system"*, EU Scoping Report. ISBN 978-92-76-43463-4. Accessible at: https://op.europa.eu/en/publication-detail/-/publication/36ebb96c-50c5-11ec-91ac-01aa75ed71a1/language-en).

Glänzel, W. (2006), *The perspective shift in bibliometrics and its consequences*. Accessible at https://de.slideshare.net/inscit2006/the-perspective-shift-in-bibliometrics-and-its-consequences

Glänzel, W., Debackere, K. (2003), *On the opportunities and limitations in using bibliometric indicators in a policy relevant context*, In: R. Ball (Ed.), Bibliometric

Analysis in Science and Research: Applications, Benefits and Limitations, Forschungszentrum Jülich (Germany), 225–236.
Glänzel, W., Schoepflin, U. (1994), Little scientometrics, big scientometrics ... and beyond? Scientometrics, 30(2–3), 375–384.
Gorraiz, J., Wieland, M., Ulrych, U., & Gumpenberger, C. (2020). De profundis: A decade of bibliometric services under scrutiny. *Evaluative informetrics: The art of metrics-based research assessment: Festschrift in honour of Henk F. Moed*, 233-260.
Gorraiz, J., Wieland, M., & Gumpenberger, C. (2016). Individual bibliometric assessment@ University of Vienna: from numbers to multidimensional profiles. arXiv preprint arXiv:1601.08049.
Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, 34(8), 575-583.
Moed, H. F., Halevi, G. (2015). Multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology*, 66(10), 1988-2002.
Moed, H. F. (2017). *Applied evaluative informetrics*. Berlin: Springer International Publishing.
Moed, H. F. (2020). Appropriate use of metrics in research assessment of autonomous academic institutions. *Scholarly assessment reports*, 2(1): 1. DOI: https://doi.org/10.29024/sar.8
Robinson-Garcia, N., Vargas-Quesada, B., Torres-Salinas, D., Chinchilla-Rodríguez, Z., & Gorraiz, J. (2024). Errors of measurement in scientometrics: classification schemes and document types in citation and publication rankings. *Scientometrics*, 1-21.
Torres-Salinas, D., Orduña-Malea, E., Delgado-Vázquez, Á., Gorraiz, J., & Arroyo-Machado, W. (2024). Foundations of Narrative Bibliometrics. *Journal of Informetrics*, 18(3), 101546.
Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. DOI: https://doi.org/10.4135/9781473978782
Wouters, P., Glänzel, W., Gläser, J., Rafols, I. (2013). The dilemmas of performance indicators of individual researchers–An urgent debate in bibliometrics. ISSI Newsletter, 9(3), 48-53

https://doi.org/10.51408/issi2025_151

The New Alliance. Bringing Together Bibliometric and Library Science for a Responsible Assessment of Research in SSH

Andrea Bonaccorsi

andrea.bonaccorsi@unipi.it
DESTEC, School of Engineering, University of Pisa, Largo Lucio Lazzarino 2, 56126 Pisa (Italy)

Abstract

We suggest a new alliance between two previously unrelated disciplines, namely Bibliometrics and Library science, with the goal of building a complete representation of the scientific production of humanities, including books and book chapters, as well as journal articles in a multilingualist perspective. We report on recent advancements on technology for interoperability of library resources that will permit an automatic validation of author identity via Authority Control. We discuss how this perspective will contribute to a fair and responsible research assessment for SSH; with particular attention to humanities.

Introduction

Among the opponents to research assessment a prominent role has been played, and is still played, by many scientific communities in humanities. Authors in fields such as history, literary criticism, or philosophy find the use of bibliometrics deeply unsatisfactory for their fields. In turn, while advocating for the use of peer review, as opposed to bibliometrics, they complain that the lack of accepted methodologies make research assessment procedures unreliable. This opinion is shared by some (not all) communities in social sciences. These arguments are well grounded.

This paper is a report on the main principles to design a new system to represent research in SSH, including some preliminary testing of the feasibility. It also includes some visionary ideas on how to use new data in order to do responsible research assessment in SSH.

The paper is organized as follows. The next section discusses the challenges to responsible research in SSH and the need for new bibliometric tools. The following section introduces the main principles and techniques to design a new data collection system. The following section introduces ideas on how to use new data in responsible evaluation of SSH. The final section concludes.

Humanities as the hidden science

While many of the issues about SSH are valid for both humanities and social sciences, they are more severe in humanities. Let us frame the discussion for humanities, and then discuss the role of social sciences at a later stage.

It has been known since long time that researchers in humanities follow a pattern of publication that differ from natural sciences (and partially differ from social sciences) (Hicks, 1999; Nederhof, 2006; Kulczycki et al. 2018). They have peculiar

information needs and practices (Stone, 1982; Watson and Boone, 1994; Wiberley, 2009; Benardou et al. 2010).

Researchers in humanities publish significantly more books than their colleagues in STEM and assign to books a higher scientific value. Existing commercial datasets (e.g. Web of Science, Scopus) do not adequately cover books and, importantly, book chapters. These bibliographic forms (that include Festschriften and proceedings of major conferences) are crucial channels for scientific communication in humanities. In addition, by design they ignore the largest part of scientific production of researchers in humanities, which takes place in national languages in non-indexed journals (Harzing et al. 2016; Federation of Finnish Learned Societies et al., 2019; Visser et al. 2021; Martín-Martin et al. 2021; Petr et al. 2021).

In STEM research is more often than not published in English to guarantee a wider circulation of the content, greater accessibility across the discipline, better ranking in search results, less opaque indexing criteria. In contrast, the language of origin is of particular importance to Humanities as it has a closer and more significant relationship with the culture in which the research is rooted. Research in humanities can face substantial obstacles if it is to avoid marginalization, particularly in very specialized areas and in non-English language research (Tsakonas, 2024). The lack of books and book chapters and the limited coverage of journals in statistics of research mean that the overall representation of humanities is enormously undervalued (Linmans, 2009). This situation has prevented the test of alternative (even conflicting) theories of citation to humanities. We just know too little. There are good reasons to believe that the patterns of STEM do not apply to humanities (Ardanuy, 2013; Hellqvist, 2010; Engels et al. 2012; Waltman, 2016; dos Santos et al. 2021). For example, citations in books are structurally different from citations in articles: they constitute a longer list, which includes more heterogeneous sources, often from a variety of disciplines, over extended time periods (Cullars, 1989; 1998). Consequently, citation analysis must be completely redefined in the case of humanities, avoiding practices such as citation count, H-index, or Impact Factor, which are common (although contested) practice in STEM.

The poor representation of humanities in data collection has deep and negative consequences in the public visibility and the impact on policymaking. This state of the art is deeply unsatisfactory.

Data on humanities is desperately missing. Researchers in philology, history, philosophy, archaeology, or literary criticism and history of art almost never show up in official statistics and in public discussions on research. They are hidden. Data on their scientific production, in particular books, book chapters and journal articles in national languages, is never on the table.

Official statistics at UNESCO, OECD or European Union level simply ignore an entire region of researchers. Nor they appear in university rankings or in the top 1% most cited authors, or the top 2% worldwide scientists. Since humanities never appear in official statistics, in the public arena of democratic societies it can be said, provocatively that they do not properly exist.

In turn, the lack of data makes it acceptable that all efforts to carry out "research on research" or even to build an ambitious "science of science" simply ignore humanities and a good part of social sciences.
In the last three decades of academic work, the word most frequently associated to "humanities" has been "crisis" (Guillory, 1993; Donoghue, 2008; Rancière, 2009; Small, 2013), even "permanent crisis". The decline in public esteem, reduction in student enrollment, cut in public funding, lack of research positions, "adjunctification" of academic careers were the most cited phenomena, in US as well as in Europe. On top of these, there is clearly a lack of self-reflection on the epistemological and methodological grounds of research in humanities. This is in sharp contrast with the high status of natural sciences. The conventional argument is that humanities do differ from natural sciences on epistemological bases. Humanities deal with indexicality, subjectivity, judgment, while natural sciences deal with regularities, objectivity, and explanation. Natural sciences produce reliable knowledge, while humanities produce opinions, implying there is no chance to put them on a par. Consequently, we currently have a fully developed science of science (see for example Wang and Barabasi, 2021) that addresses natural sciences with an ambition to move into social sciences, but we have no comparable science of science in humanities. A simple example will clarify the urgent need for going beyond the state of the art: in the multi-author article (Liu et al., 2023) that summarizes two decades of high level research in the "science of science", presumably an authoritative reference for scientific communities and policy makers alike, the word "humanities" appear only once. For those that study the way in which science is produced, *humanities do not exist*.
In recent years there have been several proposals to address the situation, mainly by leveraging on open access publications (Colavizza et al. 2023) and making use of state-of-the-art technologies for citation mining and extraction (Sula and Miller, 2014; Peroni et al. 2016; Lent et al. 2018; Colavizza et al. 2018; 2019). In particular, the proposal of a Humanities Citation Index by Colavizza et al. (2023) is remarkable. While these works have made large progress in the state of the art, several gaps still remain. First, we need to extract citations not only from open access journals but also from traditional journals, as well as (most difficult) from books and book chapters. They still form a large core that cannot be ignored. Second, we need to address the issue of Author identity, with the final goal of reconstructing the entirety of production of researchers in humanities.

Main challenges

How can we enter into a responsible framework for the assessment of SSH research? Before advancing some solutions let us review some of the most intriguing and open issues.

Book and book chapters

A well known issue is the determination of the perimeter of the book production. Let us note that this problem has been solved since long time in bibliometrics using the technique of journal indexing. On the basis of a set of formal and published criteria

the indexing organizations (Incites/Clarivate, SciVal/Elsevier or others) make a decision that defines the perimeter of analysis. On the basis of the perimeter, all kind of normalization and standardization practices can take place. *Without this technique, bibliometrics would not exist*. Now let us take note of the size of the problem with respect to books.

According to a 2023 press release from Scopus the total number of active peer-review journals is 27,950, of which 6,126 open access journals. Once the inclusion of a journal in the indexed perimeter is decided, the flow of articles (hence of authors) is automatically acquired.

It is interesting to note that the total number of books, which are not included in this flow of data, is at least one order of magnitude larger.

The organization holding the code number for books (ISBN) declares the number of books in its database as 42 million. But the definition of books and the practice of book recording from ISBN is controversial. According to an estimate by Google Books published in 2010 the number of books since the invention of print is 129,864,880. Given that UNESCO estimates that the total number of books per year is approximately 2,2 million, an updated number is 158,464,880 as for 2023. It seems that *the very definition of books is difficult to fix*, as it includes digital publishing, self-publishing, variations and repetitions that inflate the total number. A peculiar problem is also the size of Orphan books, or books for which the author is unknown or cannot be contacted. According to a study, the number of Orphan books is estimated in the order of 25 million (JISC 2009).

Given the uncertainties in the ontology of the object, as well as the large size of the universe, the goal of defining a perimeter of books, as it happens in standard bibliometrics for indexed journals, seems difficult to achieve. Shall we give up any hope? Perhaps no. Let us define the problem from a different attack point.

Authors

How many authors do exist in SSH? This question might be more manageable than the question on the number of books. To address this issue we might start with some order of magnitude from existing sources, searching for the total number of publications.

The AI-backed Dimensions, searched under the heading Human society, delivers 5,629,704 publications, of which 722,352 are book chapters, 80,917 are monographs, and 46,713 edited books. Another query on Language, Communication and Culture delivers 3,109,899 publications, of which 528,027 book chapters, 59,693 monographs, and 34,985 edited books.

Another fast-and-furious query on Open Aire delivers the following numbers: Humanities and the Arts 1,645,280 publications, Education 1,113, 256; History and archeology 721,008; Languages and literature 498,166; Philosophy, ethics and religion 316,847; Arts 111,774.

ERIH PLUS, the European Reference Index for the Humanities and Social Sciences, supported by the Norwegian Directorate for Higher Education and Skills includes more than 10,000 journals in SSH, while the number of individual authors is not declared.

The European Alliance for Social Sciences and Humanities include organizations in which 100,000 active researchers publish in SSH.

We do not know which is the share of SSH authors on the total at world level. However, there are some estimates on the total number of researchers at world level, based on UNESCO, OECD, European Commission and national data (in particular, US data). These estimate converge around a baseline number of 8 million currently active researchers worldwide (Burke et al. 2021; Ayan, Hakk and Ginther 2023). On this basis it is realistic to assume that SSH active researchers should be in the range between one and two million. If these are the numbers, the goal of building up a publicly available census of SSH authors is not out of reach, given the level of AI technologies available.

A plausible strategy might be the following. First, collect all national repositories that include only SSH authors whose scientific activity is validated. According to the survey by Sile et al. (2018) there are several European countries in which such repositories are publicly available (Kulczycki et al. 2018; 2020). This collection might create the backbone of the exercise.

Second, download authors from publicly available datasets, including Dimensions, OpenAire, Open Citations, and various repositories of open access journals. Several repositories of Open Access journals are available. The OpenDOAR directory (https://v2.sherpa.ac.uk/opendoar) already makes it possible to access to thousands of repositories across all countries. Regional federations of repositories, for example in Latin America, aggregate national and institutional repositories (e.g. https://www.lareferencia.info/es and https://www.redalyc.org/). The Directory of Open Access Books (www.doabooks.org) gives access to >80.000 books.

Third, compile an integrated list of authors with the associated metadata by integrating all publicly available sources.

Will this list be valid? Of course no. Further work should be done for the validation of authors. This problem is *not the same of disambiguation of authors in scientific journals*. The definition of author in scientific journals is very simple: any person that submits an article and gets published is ipso facto an author. The definition of the perimeter of indexed journals solves once and for all the issue of who is an author. What is left to journal publishers is the problem of disambiguation of journal authors, an issue which is largely solved by the mandatory inclusion of ORCID ID.

The largest author identification system is ORCID. The number of active records is 9,2 million in 2024, used in the same year by 2.3 billion external items (ORCID 2024) . ORCID is designed for the need of the research community and the publishing industry as a general purpose tool to reduce or mitigate the well known issue of name disambiguation. It has become the general standard, as a large number of journal editors, publishers and evaluation agencies started to ask the ORCID ID as a mandatory information for authors.

This is not the same for books and book chapters, since not all authors of books have an ORCID ID and not all authors qualify as authors of a scientific publication. Scientific publications are a subset, often a small one, of book publishing. In addition, the integration of repositories will create issues of duplication and

disambiguation. In the absence of a mandatory ORCID ID procedure, we must find another solution.

Author validation

Is there a way to establish the identity of authors without a mandatory code such as ORCID? My suggestion is that an alternative is available *in a domain of expertise that has been traditionally separated from bibliometrics*, i.e. Library science, or Information science. After extensive study of the problem and consultation with key actors, it is possible to conclude that the key is the integration between the world of libraries and the world of bibliometric datasets. There is no other way to integrate book and journal metadata in order to build up a complete representation of the scientific production of scholars in humanities. *This has never been done before*. It is a truly new alliance.

With this approach an original combination of two disciplines, previously separated, will be achieved. Library science has developed accurate methods for the disambiguation and validation of authorship, but has no interest for the aggregation of data; on the contrary bibliometrics has constructed a large array of indicators but no adequate coverage of books and book chapters, as well as of non-indexed journals, which are extremely relevant in humanities.

In this scientific and intellectual domain the issue of how to achieve a unique author identification has been crucial for decades. One can say that among the distinguished skills of authors and practitioners in libraries the correct identification of authors has traditionally been prominent, together with the methods of cataloguing.

Libraries have a robust and well tested method for the unique identification of authors, called *Authority Control*. It is defined as follows (Clack 1990, 1): "Authority control is a technical process executed on a library catalog to provide structure. Uniqueness, standardization, and linkages are the foundation of authority control".

Authority control of a library catalog is maintained through an authority file that contains the terms used as access points in the catalog. The access points that determine the structure of the catalog may be real entry headings on bibliographic records or cross references. In library catalogs the entry headings under control generally consist of personal and corporate names, uniform titles, series, and subjects.

Libraries have developed Authority Files by using over time various generations of standards and software solutions. Historically, the main problem has been the lack of interoperability of definitions and software tools. The problems are under way of solution through collaborative projects such as Share VDE (https://wiki.share-vde.org/wiki/Main_Page). This is an international library driven initiative that adopts the entity-oriented bibliographic data model BIBFRAME proposed by the US Library of Congress and the Library Reference Model defined by IFLA with the goal of making accessible bibliographic records in the Linked Open Data format (Angjeli et al. 2014; Bennett et al. 2017; Koskas, 2022; Bianchini and Sardo, 2022). Within the Share VDE project several national libraries and university libraries are currently collaborating for bringing into practice a new level of cooperation based on interoperability and openness to sustain discovery of knowledge (Possemato, 2022).

Among them it is important to mention the US Library of Congress, which has the largest global collection of books in all fields. All living authors who have published at least one book are registered.
An important implication of this collaborative effort is that it is possible (and financially plausible) *to design a software procedure for the automatic control of the Authority File* in any language and for any name of author, managing all cases of ambiguity. This is the first foundation block of the new alliance, creating a linkage between bibliometrics and library science.
Contrary to the bibliometrics based on journal indexing, the new bibliometrics will be centered around authors, whatever the entry point in the data collection system.

Citations and abstracts

At this point we might have collected an official census of authors in SSH associated with the metadata regarding books, book chapters, and articles.
The next step is to extract citation data. This exercise is largely practiced in journals but almost unknown for books and book chapters. The are two reasons: (a) citations appear in books with a variety of formats, that are not standardized (e.g. full reference in the text, full reference in the footnote, author and date in the text etc.); (b) citations include many errors, since they are self-made by authors, with limited room for an automatic control by book editors and publishers (particularly in the absence of a DOI number).
This problem is nowadays largely solvable with dedicated software that is able to automatically recognize the textual entity within the text, *using AI techniques such as Named Entity Recognition (NER)* and its more recent developments. More difficult is the problem of errors, for which limited experience is available so far. Given these hard problems, how do we address the issue of citations from books and book chapters?
The Initiative for Open Citations (www.i4oc.org) and the Initiative for Open Abstracts (www.i4oa.org) have asked publishers to deliver citations and abstracts to CrossRef, together their metadata for indexing purposes, with mixed success. It is our contention that some of the existing institutions or publishers will in the near future develop a full scale initiative to extract automatic citation data and abstract data without infringing the copyrights of publishers. There is a huge value in this enterprise, the cost of which is currently largely reduced after the advent of Large Language Models.
Let us continue my suggestions in a scenario in which fully validated citation data will be available for all authors in SSH, both citations to other works (including books) and citation from other works (including books). Abstracts will also be available in this scenario.

Acknowledgments

Let me add another desideratum. Once the software solutions for the extraction of metadata has been put in place, another opportunity will be available. Most books include a section, usually in the initial chapters (e.g. Preface, Introduction, Foreword and the like), in which authors offer a list of names of colleagues and friends who

are thanked for their collaboration with the work. While the literary style of the list is usually variable, from rigidly professional to personal and informal, the list of names offers a rich source of information. We anticipate a *new bibliometrics based on acknowledgments.*

Deceased authors

One intriguing issue in the structure of citations in books and book chapters is the *large share of citations to deceased authors*. This practice is largely different from the one in STEM, in which the life of citations is largely skewed towards recent authors (with a higher probability of citing living authors).

Citing deceased authors is a crucial practice for SSH, particularly in humanities, since the very object of study is located in the past. The epistemological role of these citations in humanities should not be underestimated (Grafton, 1999).

This however creates a serious bibliometric problem, since the computation of any citation index will be largely biased by unobserved differences in the share of deceased authors in the reference list.

Nor the problem can be addressed by discriminating the authors in the reference list using some author ID, for example ORCID. The problem with ORCID is that it is based on the principle of individual control, i.e. only authors themselves can apply for an ID and update or modify the information associated to the identity. This means that it will not be possible to build up an ORCID number for deceased authors. If we ask the FAQ system of ORCID about the ID of deceased authors the reply is the following: "Is it possible to register an ORCID iD for a deceased person?" "No. Our policy is that an ORCID iD can only be created by the individual themselves, not by any other person. This is because a core principle of ORCID is individual control. You may wish to contact ISNI (International Standard Name Identifier), as their mission is "to assign to the public name(s) of a researcher, inventor, writer, artist, performer, publisher, etc. a persistent unique identifying number"; they take a library authority approach to this, rather than a researcher-controlled one as we do".

ISNI, in turn, has 16.1 million identities for 14.3 million individual persons, of which 1.2 million are researchers (a significantly lower number than ORCID). ISNI keeps a record of deceased authors, but fails to disambiguate correctly. If we look for the record of Michael Polanyi, ISNI does not recognize that the author of *Science, faith and society* (Polanyi, 1946) is the same author of *The logic of liberty* (Polanyi, 1951). We therefore cannot rely, for different reasons, neither on ORCID nor on ISNI, irrespective of their respective values and contributions.

Within the proposed new alliance it is possible to refer again to the Authority Control methodology. Authority Files, as opposed to ORCID files, include the dates of publication of all works by the same author, *even if he/she deceased*. As opposed to ISNI identities, there are no errors or ambiguity. Using some conventions on the latest dates of publication we might identify deceased authors with reasonable approximation.

An automatic procedure might therefore *classify all citations in two categories of active vs. deceased authors* and calculate the citation indexes separately. The

classification might be updated dynamically at regular intervals to take into account changes in the proportion between the two categories.

Academic publishers

While the Authority Control made possible by the library system will eliminate ambiguity on author identities, it will not per se discriminate with respect to the scientific content of the publication. This issue might be complicated by the circumstance that many academic authors do publish academic works alongside popular science publications, or collection of newspaper articles and book reviews.
While the general issue might remain controversial, a practical solution might be to refer to the list of academic publishers established by the Spanish CSIC (Gimenez-Toledo et al. 2019).

Affiliation

This information will be generally available in the metadata from journals and books, but several problems must be addressed. A combination of methods should be used here: official registers and AI.
First, it is extremely likely that the metadata will include definitions of the affiliation that are not standardized. It will be possible to use the available standard definitions of affiliations, such as ISNI (www.isni.org), ETER (European Tertiary Education Register) for European higher education institutions (https://eter-project.com/) and the ORGREG register for Public Research Organizations (https://www.risis2.eu/registers-orgreg/). Non-European affiliations will be checked against UNESCO datasets (https://www.whed.net/home.php).
Second, it is possible that in some cases the metadata on affiliation will be missing. In this case an AI-backed procedure will search for affiliation data of the identified author associated to dates and might produce an estimate of the affiliation for the missing publication.

The strenght of the new alliance

The strenght of the proposal lies in the alliance between bibliometrics and library science. The automatic validation of authors using Authority Files will ensure that all data, whatever the source of collection, will land into a validated database.
In turn, the classification of cited authors by age (in particular, the discrimination between living or recent authors and deceased authors) *will allow the deployment of the bibliometric toolbox* with respect to standardization and normalization of data.
This will create an incentive for publishers to deliver their metadata (including citations and abstracts) on a regular basis, in order to fill the census with their own data. Remaining outside the platform will be too costly. The idea needs someone who makes the initial investment and opens the way.

From the new alliance to responsible research assessment

The new alliance between bibliometrics and library science might deliver solutions that made it possible to improve the quality of research assessment and address the issues of transparency, diversity and fairness. Let us articulate this proposition.

It is fair to say that the dominant methodology for research assessment in SSH is the peer review. We know from a large literature, however, that peer review is not the golden age of research assessment. It has its own methodological weaknesses and is subject to biases of various types.

We need to go beyond the notion of informed peer review, whereby the individual peer review is assisted by a few simple bibliometric data such as citation count or citation weight. Let us consider a scenario in which quantitative bibliometrics will deliver qualitative insights that support and complement human evaluation.

In other words it is possible to anticipate a scenario in which

- a census of validated authors in SSH is established
- for each of the works of validated authors we have metadata
- metadata include citations, ackowledgments and abstracts
- data is available based on formats that allow large scale processing.

In this scenario we might give full justice to the humanities by addressing, first of all, the controversial issue of productivity. It is often assumed that research in humanities is less cumulative and less convergent than in STEM, hence less productive (Cole 1983; 1994; Clauset et al. 2015). The issue is controversial (Hedges 1987; Fawcett and Higginson 2012; Fanelli and Glänzel, 2013). A few years ago *Nature* made the claim that humanities, or soft science, should be preserved and protected (Nature 2015), but the issue of relative productivity has never been addressed systematically.

Are researchers in humanities less productive? No analysis of productivity can be done without the definition of the perimeter of the overall scientific production. To the extent that data collection is successful we might address several open (and contested) issues. Does the scientific production of researchers in humanities follow the same skewed distribution that we find in natural sciences? Is it subject to the Matthew effect? Does it decline with age or academic age? Is it associated to academic position, affiliation, type of institution? What is the typical life cycle of scientific production? On all these issues the current evidence is limited and scattered. Recent research has shown that researchers in humanities do not differ from STEM in the shape and asymmetry of the distribution of scientific production, following the so called Matthew effect (Bonaccorsi et al. 2017). A related issue is whether researchers in humanities adopt team production and authorship. Are researchers adopting the team-based inquiry approach of their colleagues in STEM? In which disciplines do we find a larger average (and median) number of co-authors? Does the size of team has an influence on the degree of novelty produced (Wu et al. 2019)?

In this scenario a whole range of Natural Language Processing techniques can be introduced, tested and validated as a support to human judgment. They might be a powerful support to responsible assessment of research. They include embeddings

and variable length embeddings, network dynamics, knowledge graph, sentiment analysis, citation networks, citation clustering and many others (Chen et al. 2009; Guevara et al. 2016; Kozlovski et al 2018; Chinazzi et al. 2019; Tshitoyan et al. 2019; Miao et al. 2022; Peng et al. 2021).

Scientific texts are an optimal field for data analysis, because researchers speak a controlled language that is, by design, aimed at being critically evaluated. Recent technologies in NLP and pre-trained LLM systems allow a fine-grained analysis of the content of scientific publications, with unprecedented sophistication.

Thus for each of the main (and controversial) issues in the epistemology of humanities it will be possible implement one or more AI-based technique: word embeddings to examine the novelty of knowledge produced by humanities; Knowledge Graphs to examine the explanatory nature of statements and the cause-effect relations; Topic Modeling and citation clustering to study the formation of scientific consensus, the persistence of paradigmatic pluralism and the management of controversies; citation networks and field tracking to investigate into the cumulativeness of knowledge; again embeddings, but also information density and linguistic complexity to explore the level of interdisciplinarity.

Topic modeling (as in Bonaccorsi et al. 2022) and word embeddings (as in Melluso et al. 2024a; 2024b) might be applied to the collection of books and articles described above. Recently developed methodologies in the full text processing of publications, such as information density (Bernstein, 1964; Bischhof and Eppler, 2010; Evans and Aceves, 2016; Hamilton et al. 2016; Aceves and Evans, 2023) and linguistic complexity (Lu et al. 2019a; 2019b) allow a granular analysis of the structure of argumentation. The extent to which they can be replicated on abstracts is to be explored.

Conclusions

This paper suggests a new alliance between bibliometrics and library science in order to build up a responsible assessment for SSH. The evaluation of research in these fields requires the full scale consideration of books, book chapters, and journal articles in a multilingualist perspective.

My proposal is complementary to the institutional efforts, undertaken by the European Union, to establish Open Science, through the creation of a European Quality Standard for Institutional Open Access Publishing (EQSIP) (e.g. https://diamasproject.eu), the technical improvements of open journal platforms for the Diamond OA (www.craft-oa.eu) and the exploration of open metric data such as OpenCitations (https://opencitations.net) and Scholexplorer (https://scholexplorer.openaire.eu). With respect to these efforts one of the major limitations is that books and book chapters have very limited coverage in open access

and even, as addressed by the Palomera project (https://operas-eu.org/projects/palomera/) in open access funding.

This paper argues that the technological resources to undertake the enterprise of a new alliance are available.

References

Aceves, P., Evans, J.A. (2023) Human languages with greater information density increase communication speed, but decrease conversation breadth. Pre-print.

Angjeli, A., MacEwan, A., Boulet, V. (2014) ISNT and VIAF. Transforming ways of trustfully consolidating identities. IFLA WLIC 2014 Conference. Lyon, 1-19.

Ardanuy, J (2013) Sixty years of citation analysis studies in the humanities (1951–2010). Journal of the American Society of Information Science and Technology, 64(8), 1751–1755.

Ayan, D.E., Hakk, L.L., Ginther, D.K. (2023) How many people in the world do research and development? Global Policy. DOI: 10.1111/1758-5899.13182

Benardou, A., Constantopoulos, P., Dallas, C., Gavrilis, D. (2010) Understanding the information requirements of arts and humanities scholarship. International Journal of Digital Curation 5(1), 18–33.

Bennett, R., Helgel-Dittrich, C., O'Neill, E.T.O., Tillett, B.B. (2017) VIAF (Virtual International Authority File): Linking the Deutche Nationalbibliothek and Library of Congress Name Authority Files. International Cataloguing and Bibliographic Control, XXXVI, 1, 12-18.

Bernstein, B. (1964) Elaborated and restricted codes. Their social origins and some consequences. American Anthropologist, 66, 55-69.

Bianchini, C., Sardo, L. (2022) Wikidata: A new perspective towards universal bibliographic control. In G. Bergamin and M. Guerrini (eds.) Bibliographic control in the digital ecosystem. Florence, AIB-EUM-Firenze University Press.

Bischhof, N., Eppler, M.J. (2010) Clarity in knowledge communication. Proceedings of the Tenth International Knowledge Management Conference Iknow, vol. 10, 162-174.

Bonaccorsi A. (2018) Towards an Epistemic Approach to Evaluation in SSH. In Bonaccorsi A. (ed.) (2018) The evaluation of research in Social Sciences and Humanities. Lessons from the Italian experience. New York, Springer International Publishing.

Bonaccorsi A., Daraio C., Fantoni S., Folli V., Leonetti M., Ruocco G. (2017) Do Social Sciences and Humanities behave like life and hard sciences? Scientometrics, 112, 607-653.

Bonaccorsi, A. (2023) An epistemic approach to research assessment in the social sciences. In Tim C.E. Engels, Emanuel Kulczycki (eds.) Handbook on research assessment in the Social Sciences. Cheltenham, Edward Elgar.

Bonaccorsi, A. (2025) The knowledge of humanities. A comparative epistemology of historiography, literary criticism, history of art, and history of architecture. Turnhout, Brepols (forthcoming).

Bonaccorsi, A., Melluso, N., Massucci, A. (2022) Exploring the antecedents of interdisciplinarity at the European Research Council: a topic modeling approach. Scientometrics. https://doi.org/10.1007/s11192-022-04368-9

Burke A, Finamore J, Foley D, Jankowski J, Moris F; National Center for Science and Engineering Statistics (NCSES) (2021). Measuring R&D Workers Using NCSES Statistics. NSF 21-335. Alexandria, VA: National Science Foundation. Available at https://ncses.nsf.gov/pubs/nsf21335/.

Chen, C., Chen, Y., Hou, H., Liu, Z., Pellegrino, D. (2009) Towards an explanatory and computational theory of scientific discovery. Journal of Informetrics, 3, 191-209.

Chinazzi, M., Gonçalves, B., Zhang, Q., Vespignani, A. (2019) Mapping the physics research space. A machine learning approach. APJ Data Science, 8, 33.
Clack, D.H. (1990) Authority Control: Principles, Applications, and Instructions. Chicago, American Library Association.
Clauset A., Arbersman, S., Larremore, D.B. (2015) Systematic inequality and hierarchy in faculty hiring networks. Science Advances 1, e1400005.1
Colavizza, G., Peroni, S., Romanello, M. (2023) The case for the Humanities Citation Index (HuCI): A citation index by the humanities, for the humanities. International Journal on Digital Libraries, 24, 191-204.
Colavizza, G., Romanello, M., Babetto, M., Barbay, V., Bolli L., Ferronato, S., Kaplan, F. (2018) Linked Books: Towards A Collaborative Citation Index for the Arts and Humanities. Proceedings of the Red de Humanidades Digitales, A. C., Mexico City, 178–181.
Colavizza, G., Romanello, M. (2019) Citation mining of humanities journals: the progress to date and the challenges ahead. Journal of European Periodical Studies, 4, 36–53.
Cole, S. (1983) The hierarchy of the sciences? American Journal of Sociology, 89, 111-139.
Cole, S. (1994) Why sociology doesn't make progress like the natural sciences. Sociological Forum, 9, 133-154.
Cullars, J. M. (1989). Citation characteristics of French and German literary monographs. Library Quarterly, 59 (305–325).
Cullars, J. M. (1998). Citation characteristics of English-language monographs in philosophy. Library and Information Science Research, 20(1), 41–68.
Donoghue, F. (2008) The last professors. The corporate university and the fate of the humanities. New York, Fordham University Press.
dos Santos, E.A., Peroni, S., Mucheroni, M.L.(2021) Citing and referencing habits in medicine and social sciences journals in 2019. Journal of Documentation. 77(6), 1321–1342.
Engels, T. C., Ossenblok, T. L., & Spruyt, E. H. (2012). Changing publication patterns in the social sciences and humanities, 2000–2009. Scientometrics, 93(2), 373–390.
Evans, J.A, Aceves, P. (2016) Machine translation. Mining text for social theory. Annual Review of Sociology, 42(1), 21-50.
Fanelli, D., Glänzel, W. (2013) Bibliometric evidence for a hierarchy of sciences. PLoS ONE 8, e66938.
Fawcett, T.W., Higginson, A.D. (2012) Heavy use of equations impedes communication among biologists. PNAS, 109, 11735-11779.
Federation Of Finnish Learned Societies, Information, T.C.F.P., Publishing, T.F.A.F.S., Universities Norway, ENRESSH (2019) Helsinki initiative on multilingualism in scholarly communication. Figshare.
Fodor, J.A. (1974) Special sciences (or: the disunity of science as a working hypothesis). Synthese, 28(2), 97-115.
Fortunato, S., Bergstrom, C.T., Borner, K., Evans, J.A., Helbing, D., Milojevic, S., Petersen, A.M. et al. (2018) Science of science. Science, 359(6379), eaao0185.
Giménez-Toledo, E. Sivertsen, G., Mañana-Rodriguez, J. (2019) International Register of Academic Book Publishers (IRAP): overview, current state and future challenges. In Daraio et al. (eds.) Proceedings of the S&T Indicators Conference. Rome, Efesto Publishers.
Grafton, A. (1999) The footnote: A curious history. Cambridge, Mass. Harvard University Press.
Guevara, M.R., Hartman, D., Aristarán, M., Mendoza, M., Hidalgo, C.A. (2016) The research space: Using career paths to predict the evolution of the research output of individuals, institutions, and nations. Scientometrics, 109, 1695-1709.

Guillory, J. (1993) Cultural capital. The problem of literary canon formation. Chicago, University of Chicago Press.
Hamilton, W.L., Leskovec, J., Jurafsky, D. (2016) Diachronic word embeddings reveal statistical laws of semantic change. arXiv/org/abs/1605.09096.
Harzing, A.-W., Alakangas, S. (2016) Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. Scientometrics 106(2), 787–804.
Hedges, L.V. (1987) How hard is hard science, how soft is soft science? The empirical cumulativeness of research. American Psychologist, 42, 443-455.
Helbing, D. (2012) Accelerating scientific discovery by formulating grand scientific challenges. The European Physical Journal Special Topics, 214(1), 41-48.
Hellqvist, B. (2010) Referencing in the humanities and its implications for citation analysis. Journal of the American Society of Information Science and Technology, 61(2), 310–318
Hicks, D. (1999) The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. Scientometrics 44(2), 193–215.
Koskas, M. (2022) Universal bibliographic control today: Preliminary remarks. In G. Bergamin and M. Guerrini (eds.) Bibliographic control in the digital ecosystem. Florence, AIB-EUM-Firenze University Press.
Kozlowski, A.C., Taddy, M., Evans, J.A. (2018) The geometry of culture. Analyzing meaning through word embedding. American Sociological Review, 84, 905-949.
Kulczycki, E., Engels, T.C.E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Isteni-Stari, A., Zuccala, A. (2018) Publication patterns in the social sciences and humanities: evidence from eight European countries. Scientometrics, 116(1), 463–486.
Kulczycki, E., Guns, R., Pölönen J., … Sivertsen, G. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. Journal of the Association for Information Science and Technology. 1–15. DOI:https://doi.org/10.1002/asi.24336
JISC (2009) In from the Cold An assessment of the scope of 'Orphan Works' and its impact on the delivery of services to the public. Available at https://web.archive.org/web/20091118103903/http://sca.jiscinvolve.org/files/2009/06/sca_colltrust_orphan_works_v1-final.pdf#
Lent, H., Hahn-Powell,G., Haug-Baltzell, A.,Davey, S., Surdeanu, M., Lyons, E. (2018) Science citation knowledge extractor. Frontiers of Research Metrics and Analysis, 3, 35.
Linmans, A.J.M. (2009) Why with bibliometrics the humanities does not need to be the weakest link: indicators for research evaluation based on citations, library holdings, and productivity measures. Scientometrics, 83(2), 337–354.
Liu, L., Jones, B.F., Uzzi, B., Wang, D. (2023) Data, measurement and empirical methods in the science of science. Nature Human Behaviour, 7, 1046-1058.
Lu, C. et al. (2019a) Analyzing linguistic complexity and scientific impact. Journal of Informetrics, 13, 817-829.
Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schaars, M., Zhang, C. (2019b) Examining scientific writing styles from the perspective of linguistic complexity. Journal of the American Society of Information Science and Technology, 70(5), 462-475.
Martín-Martín, A., Thelwall, M., Orduna-Malea, E., Delgado López-Cózar, E. (2021) Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. Scientometrics, 126(1), 871–906.
Melluso, N., Bonaccorsi, A. (2024b) Novelty and interdisciplinarity. Mimeo.

Miao, L., Murray, D., Jung, W.S., Larivière, V., Sugimoto, C.R., Ann, Y.Y. (2022) The latent structure of global scientific development. Nature Human Behavior, 6, 1206-1217.
Nature (2005) In praise of soft science. Nature Editorial, 435, 1003.
Nederhof, A.J. (2006) Bibliometric monitoring of research performance in the social sciences and the humanities: a review. Scientometrics, 66(1), 81–100.
Peng, H., Ke, Q., Budek, C., Romero, D.M., Ann, Y.Y. (2021) Neural embeddings of scholarly periodicals reveal complex disciplinary organizations. Science Advances, 7(17), eabb9004.
Peroni, S., Shotton, D., Vitali, F. (2016) A document-inspired way for tracking changes of RDF data. Proceedings of the 1stWorkshop on Detection, Representation and Management of Concept Drift in Linked Open Data. CEUR Workshop Proceedings-
Petr, M., Engels, T.C.E., Kulczycki, E., Dušková, M., Guns, R., Sieberová, M., Sivertsen, G. (2021) Journal article publishing in the social sciences and humanities: a comparison ofWeb of Science coverage for five European countries. PLoS ONE 16(4), 0249879.
Possemato, T. (2022) Entity modelling. JLIS.it, XIII, 3, 12-28.
Rancière, J. (2009) Aesthetics and its discontents. Cambridge, Polity Press.
Sīle, L., Pölönen, J., Sivertsen, G… Teitelbaump, R. (2018). Comprehensiveness of national bibliographic databases for social sciences and humanities: findings from a European survey. Research Evaluation, 27(4), 310–322. DOI: https://doi.org/10.1093/reseval/rvy016
Small, H. (2013) The value of the humanities. Oxford, Oxford University Press.
Stone, S. (1982) Humanities scholars: information needs and uses. Journal of Documentation, 38(4), 292–313.
Sula, C.A., Miller, M. (2014) Citations, contexts, and humanistic discourse: toward automatic extraction and classification. Literary and Linguistic Computing, 29(3), 452–464.
Tsakonas, G. (2024) Big cultures and small languages: A new paradoxography in a shifting research system. Paper presented to the Fiesole Retreat Conference (available at https://youtu.be/_15zEKBKSM4).
Tshitoyan, M.L., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O. et al. (2019) Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 571, 95-98.
Visser, M., van Eck, N.J., Waltman, L. (2021) Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies. 2(1), 20–41
Waltman, L. (2016) A review of the literature on citation impact indicators. Journal of Informetrics, 10(2), 365–391.
Wang, D., Barabàsi, A.L. (2021) The science of science. Cambridge, Cambridge University Press.
Watson-Boone, R. (1994) The information needs and habits of humanities scholars. RQ 34(2), 203–215.
Wiberley, S.E., Jr. (2009) Humanities literatures and their users. Encyclopedia of Library and Information Sciences, 3rd edition, 2197–2204.
Wu, L., Wang, D., Evans, J.A. (2019) Large teams develop and small teams disrupt science and technology. Nature, 566 (7744), 378-382.

https://doi.org/10.51408/issi2025_152

Trueblood et al.'s Ideas on Research Evaluation and Implications for Reforming Research Assessment

Ronald Rousseau

ronald.rousseau@uantwerpen.be, ronald.rousseau@kuleuven.be
University of Antwerp, Faculty of Social Sciences, 2020 Antwerp (Belgium)
KU Leuven, MSI, Facultair Onderzoekscentrum ECOOM, 3000 Leuven (Belgium)

Abstract

This contribution to the FRAME track at ISSI 2025 offers a brief overview of Trueblood et al. (2025), highlighting its relevance for research evaluation. In their article, Trueblood and 14 co-authors examine the current publication landscape and explore both how it can be transformed and how such changes logically necessitate a shift in research assessment practices.

Introduction

Early this year, Trueblood and 14 co-authors published "The misalignment of incentives in academic publishing and implications for journal reform" in the Proceedings of the National Academy of Sciences of the USA (PNAS). The authors studied the publishing landscape, decry how commercial publishing companies generate huge profits, and propose ways to let academic institutions regain control of scientific publishing. As scientific publications play an important role in research evaluations, the authors spend several pages on reforms in academic evaluation. In this contribution to the FRAME track at ISSI 2025, a short overview of Trueblood et al. (2025) is presented with an emphasis on its implications for research evaluation. In this document the expression "the authors" is used for Trueblood et al. (2025).
The authors noted that the two main goals of publishing, namely the documentation of new knowledge and establishing scientific credentials are often in tension. It is, indeed, well-known that even in the best of circumstances, maximizing metrics may lead scientists to prioritize novelty and even sensationalize findings to publish in so-called prestigious journals. In this way, important details and partial null results may be hidden from view.

Academic publishing: now and in the past

The development of academic publishing is closely connected to the growth of universities, the formation of scientific societies, and the professionalization of academia. Trueblood et al. (2025) write that in 1950 there were about 10,000 journals worldwide. This number increased to 62,000 in 1980 and according to (Suiter and Sarli, 2019) to 80,000 in 2019. Note that in 2015 an estimate for the number of Chinese journals has been published, reaching a total of at least 8,000 journals (Rousseau, 2015). Commercial publishing companies such as Elsevier, Springer Nature, Wiley-Blackwell, and Taylor & Francis, dominate a large part of modern scientific publishing. It is sometimes argued (Fyfe et al., 2017) that these firms

exploit reviewers and editorial boards by requiring free services, making it costly to distribute scientific work, and levying high fees for open access.
Trueblood et al. (2025) further discuss the role of modern journals. They offer details from three perspectives: 1) journals as revenue streams, including the bad sides of it such as predatory journals and paper mills; 2) journals as curators of research and the role of peer review; 3) journals as the cornerstone of the academic prestige economy, leading to problems such as "publish or perish" and the pressure to continuously produce and publish scholarly work, preferably in high-prestige journals.

Alternative publishing models

The authors argue that academic institutions and learned societies should take over the journal publishing industry, turning it into a nonprofit sector, where science controls science. New publishing models, including preprint platforms, must be established. They propose some ideas and offer examples of existing initiatives.
- Academia retaking control
As commercial publishers are not likely to stop their business or hand it over to academia, the following steps are proposed. First, academic institutions and associations create new journals controlled by themselves and ask editorial teams now working for for-profit journals to switch to the new journals. Second, scientific academies and societies should support this switch and ask their members to stop working for commercial companies, either as editors, reviewers or as authors. Third, as academics cannot take care of the purely technical work of journal publishing, competitive calls must be made so that experienced companies can do this work at lower costs. Note that this comes close to what ISSI has done when switching from Elsevier to MIT Press. In conclusion: publishing must be by scientists for scientists.
- Preprint servers
In certain fields preprints and society proceedings are considered already as more reputable than journal publications. This has happened because highly cited articles may reside solely on preprint servers. Perelman's Fields Medal winning mathematical work has never been formally published (he even refused the Medal). In our field we have (Larivière et al., 2016), cited 220 according to Google Scholar, but never formally published in a journal. Generally, and in all fields, preprints are becoming increasingly valuable. Besides citation counts, also download counts are calculated for preprint papers.
- Journal reviewed preprints
This section answers the problem that preprint servers contain papers that are not peer-reviewed. Here Trueblood et al. (2025) refer to the new policy adopted by eLife. Nowadays this journal only reviews articles made available on a dedicated preprint server. The outcome of these reviews is no longer used as a basis for an accept/reject decision, and the number of articles hosted on eLife is not limited. The decision of whether to host a submission, before review, is made by scientifically active editors, and reviews are presented as commentaries alongside the article. Although the authors consider this approach a significant change, they do not call it "disruptive", as editors still determine which papers will be hosted on the preprint server. The idea

of journal reviewed preprints can focus the publishing process on improving reporting and facilitating knowledge diffusion.
Yet, because the method differs so much from the traditional way, Clarivate has decided that eLife will not receive an impact factor in 2025.
- Community reviewed preprints
PCI (Per Community In) also undertakes to review preprints. PCI is a community-sourced service that provides free, journal-independent reviews of preprints. Preprint authors and reviewers collaborate to improve the preprint, ultimately leading to a recommendation where a recommender (serving a similar role as a journal editor) endorses the article for publication. The process may conclude when a PCI-recommended preprint is published on the corresponding thematic PCI websites with a DOI, allowing it to be cited, or it may be published directly in a PCI-friendly journal. PCI and similar initiatives align perfectly with an open access framework. The concept of community-reviewed preprints bears similarities to the idea proposed by Perakakis et al. (2010) under the name of Natural Selection of Academic Papers. Instead of a service like PCI, they envisioned a Global Open Archive containing the original preprint (possibly via an institutional preprint server), open and signed reviews (including those initiated by the authors), updates, and citations. In this open environment, also reviewers could be rated.
- Society endorsed preprints
The role of PCI could be assumed by societies, enlarging the prestige of this approach. The authors suggest that also ArXiv, and similar preprint servers, could play this role. They further propose that federal granting agencies and private foundations could supply the resources needed to support these changes. However, they note that this approach still carries the notion of prestige, much like traditional journals. As a result, they conclude the section by suggesting that perhaps scientists should move away from using publications as a measure of prestige.
- Modular publishing platforms and micropublications
Modular publishing breaks up a paper into small sections called modules. According to the authors F1000, eLife and PLOS Biology already publish micropublications, small articles without a broader context. Some platforms such as Octopus allow the threading of modules into a coherent narrative.

Barriers to change

The more new publishing models differ from the classical way publications are handled, the more difficult it is to become broadly accepted. There is a monetary cost and the cost of extra learning and effort, deterring potential adopters.
Pay-to-publish is an obvious choice for these new models, but requires that scientists have funds at their disposal and micropublications do not seem to correspond with the way some fields, such as the humanities see scholarly work.
Trueblood et al (2025) end this part with reflections on barriers to change by stating the following three big challenges to journal reform:

a) The lack of independence of most scientific journals from commercial for-profit publishing companies

b) The financial impacts on societies that nowadays generate substantial revenue from their journals

c) Resistance to adoption because of concerns regarding academic prestige

This leads to the topic of this conference section.

Reforms in academic evaluation

For most scientists today, publishing needs to translate into career value, namely, recognition by hiring committees and funding agencies. Therefore, the reputation of a publication venue is crucial. This highlights the need to reform academia's incentive structure, but the authors caution that such changes could bring unintended consequences. They point out that the current system has already contributed to the rise of paper mills and so-called 'predatory journals'. To better understand these potential pitfalls, they suggest that applying game theory could offer valuable theoretical insights.

Next, the authors consider five ways in which to reform academic evaluation.

Abandoning problematic metrics

They recall that citation counts and journal impact factors are highly problematic. The number of received citations depends on many factors, many of which are independent of research quality (including pure luck). Since Seglen (1989, 1997) and its replications (Zhang et al., 2017), researchers know that impact factors should not be used to judge papers. This insight has led to many reactions from the scientific and publishing community such as, e.g., DORA (2012). Moreover, altmetrics are even more easily gamed than citations. As examples of positive evolutions, the authors mention the introduction of narratives and evidence-based curriculum vitae. Of course, an academic career path does not only include published articles, but also books, teaching, and outreach activities.

Adopting responsible metrics

Here the authors mention the Leiden Manifesto for Research Metrics (Hicks et al., 2015), emphasizing best practices and allowing researchers to hold their evaluators to account. They further recall that using evidence-based CVs is an example of a best practice, as it leads to transparency. Yet, one must recognize that there are at the moment systemic issues in the evaluation process.

Quantitative metrics: measuring researcher impact

In this section, Trueblood et al. (2025) consider the evaluation of researchers and discuss different ways of weighing authors' contributions. They do not come to a concrete proposal for this very tricky problem. In some counting systems, for an overview of counting systems we refer to (Gauffriau,2021) adding authors decreases the score of the others, which could result in scientists with disadvantaged backgrounds being relegated to the acknowledgment section.

Quantitative metrics: counting replications

The push for novelty makes researchers reluctant to try to replicate others' work. Yet, it is well-known that many published results cannot be replicated, the so-called "replication crisis". It is suggested that the number of replications (and I add also direct extensions) could be a measure of interest created in the field (or even outside). If successful, replications, and replications of replications can be a measure of the robustness of the original research. Note that the authors even include unsuccessful replications, of course not in terms of robustness but in terms of research interest. They, correctly, warn that emphasis on replications and reproducibility should not divert scarce resources.

Rewarding societal impact

Since much of the research is funded by public institutions, it's reasonable for the public to expect some return on that investment. In this context, scientists are expected to engage with the broader community and, ideally, to make a visible impact. While measuring the outcomes of such engagement can be challenging, a useful starting point might be tracking the input, namely, how often scientists interact with the public. Talking about the societal impact the authors refer to Overton.io for the quantification of the policy influence of publications and of grey literature such as technical reports.

Incentivizing quality over quantity

The authors point out that there are some easy and relatively minor changes possible in academic evaluation to alter too narrow incentive structures. Focusing on top x publications in the latest y years is such a simple measure. This alters the focus from quantity to quality. Of course, it is supposed that evaluators actually read these papers, otherwise, the focus would shift again to "high-level journals", or worse to impact factors. The authors provide examples of funding organizations that take this approach such as ERC (Europe) and the NSF in the USA. Papers must be separated from the journals in which they are published. This is a must when papers are not published in the traditional sense of the word, cf. the earlier section on publishing.

Discussion and Conclusions

The influence of commercial publishers and the academic prestige economy have both a detrimental influence on scientific quality and the idea of science for the benefit of humankind. Biased incentives have even led to academic fraud such as using paper mills to increase the number of publications and to other fraudulent behavior.

The authors propose that publishing goals should be aligned with the broader aims of knowledge creation and dissemination. In this spirit, they suggest several alternative publishing approaches and encourage the scientific community to explore and incorporate these into research assessment practices. They acknowledge, however, that metrics based on time and effort are inherently more complex and harder to interpret than those based on straightforward counts.

In conclusion, Trueblood et al. (2025) urge the research community to reshape academic publishing to better serve researchers and academia. Reforming the landscape of scientific publishing naturally leads to implications for research assessment.

References

DORA (2012). San Francisco declaration on research assessment. DORA – ASCB. Available from http://www.ascb.org/dora

Fyfe, A., Coate, K., Curry, S., Lawson, S., Moxham, N., & Røstvik, C.M. (2017). *Untangling academic publishing: A history of the relationship between commercial interests, academic prestige and the circulation of research*. Discussion Paper. University of St Andrews. https://eprints.bbk.ac.uk/id/eprint/19148/. Accessed April 11, 2025.

Gauffriau, M. (2021). Counting methods introduced into the bibliometric research literature 1970–2018: A review. *Quantitative Science Studies*, 2(3), 932–975.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). The Leiden Manifesto for research metrics. *Nature*, 520(7548), 429-431.

Larivière, V., Kiermer, V., MacCallum, C.J., McNutt, M., Patterson, M., Pulverer, B., Swaminathan, S., Taylor, S., & Curry, S. (2016). A simple proposal for the publication of journal citation distributions. bioRxiv preprint doi: https://doi.org/10.1101/062109

Perakakis, P., Taylor, M., Mazza, M., & Trachana, V. (2010). Natural selection of academic papers. *Scientometrics*, 85(2), 553–559.

Rousseau, R. (2015). The tip of the Chinese publication iceberg. *ISSI Newsletter* #44, 11(4), 100-102.

Seglen, P.O. (1989). From bad to worse: Evaluation by journal impact. *Trends in Biochemical Sciences*, 14(8), 326-327.

Seglen, P.O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*, 314(7079), 498-502.

Suiter, A.M., & Sarli, C.C. (2019). Selecting a journal for publication: Criteria to consider. *Missouri Medicine*, 116(6), 461-465.

Trueblood, J.S., Allison, D.B., Field, S.M., Fishbach, A., Gaillard, S.D., Gigerenzer, G., Holmes, W.H., Lewandowsky, S., Matzke, D., Murphy, S.D., Musslick, S., popov, V., Roskies, A.L., ter Schure, J., & Teodorescu, A.R. (2025). The misalignment of incentives in academic publishing and implications for journal reform. *Proceedings of the National Academy of Sciences of the United States of America*, 122(5), e2401231121.

Zhang, L., Rousseau, R., & Sivertsen, G. (2017). Science deserves to be judged by its contents, not by its wrapping: Revisiting Seglen's work on journal impact and research evaluation. *PLoS ONE*, 12(3): e0174205.

https://doi.org/10.51408/issi2025_153

Responsible Research Assessment of Teams: Reflections and Perspectives After Two Evaluation Cycles at the University of Antwerp, Belgium

Tim C.E. Engels[1], Birgit Houben[2], Pieter Spooren[3]

[1]*tim.engels@uantwerpen.be*
Centre for R&D Monitoring (ECOOM), Faculty of Social Sciences, and Department of Research, Innovation & Valorisation Antwerp (RIVA), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

[2]*birgit.houben@uantwerpen.be*
Department of Research, Innovation & Valorisation Antwerp (RIVA), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

[3]*pieter.spooren@uantwerpen.be*
Department of Research, Innovation & Valorisation Antwerp (RIVA), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

Abstract

The University of Antwerp started research assessments per discipline that include a site visit by an international panel of peers in 2007. A few years later we reported that for research teams in the sciences basic metrics like group size, h-index and efficiency in publishing in top journals predicted panel assessments of quality and productivity (Engels et al, 2013). Upon the completion of the second cycle of research assessments in the current academic year 2024-2025, we ask ourselves (1) to what extent the stated aim of improving the quality and impact of research has been achieved, and (2) what shape the third cycle of research assessments will take. For this third cycle, the need to reconcile quantitative and qualitative approaches, responsible use of metrics, transparency and inclusivity are top priorities.

In this paper we first analyze and reflect upon the evolution, the results and the lived experiences of the UAntwerp research assessments since 2007. We then present our proposal for the third cycle of UAntwerp research assessments that will focus on creating contexts in which research can flourish. To achieve this, the assessments will take achievements and bibliometric and other indicators as context elements rather than as elements of assessment. Our aim is to launch a system of more responsible research assessments that will be fully formative and future-oriented, with validated dashboards capturing inputs, process elements, outputs, and impacts as context elements for qualitative assessments.

Introduction

Research assessment needs to consider input, process, output and impact of research, whereby impact involves both scholarly-academic impact as well as broader cultural, economic and social impact (Moed, 2017). In practice, however, research assessment too often mainly relates to bibliometric indicators of journal articles indexed in citation indexes such as Web of Science or Scopus. Even though bibliometricians have repeatedly stressed the important limitations of the use of bibliometrics when assessing individual researchers (e.g. Wouters et al., 2013), it seems that the omnipresence of bibliometric indicators has taken over research assessment at many

levels (Wilsdon et al., 2015), leading to calls to seriously rethink their use (CoARA, 2022; Zhang & Sivertsen, 2020), as well as fierce debate about bibliometrics versus peer review (Abramo, 2024).

In our modest opinion, a debate that is just as important is how to conceive responsible research assessments that do consider input, process, output and impact of research, thereby integrating qualitative and quantitative approaches. Since research assessment and research behaviour co-evolve (OECD Global Science Forum, 2025), and science has become a team effort in a majority of cases, there is a pressing need to rethink research assessment of teams. Such research assessments should welcome inclusive research and a diversity of outputs and impacts, while emphasizing the importance of a research context, research environment, and research process conducive to responsible research and innovation with impact. In this paper we explore, after two cycles of research assessments of teams at the University of Antwerp, how the next cycle of research assessments at our university can be brought more in line with the ambitions of the responsible research assessment agenda (Global Research Council, 2024).

Investments in and evaluation of university research in Flanders

According to the regional innovation scoreboard of the European Commission, Flanders is an innovation leader. The region scores particularly well in terms of international scientific co-publications and public-private co-publications, illustrating the large extent of internationalization of university research and the strong integration of innovation ecosystems in the region. Although government expenditures on R&D remain well below 1% of the regional GDP, business expenditures on R&D have increased significantly over the last decade and are currently well above 3% (IDEA Consult, 2024) As for Europe as a whole, boosting productivity and competitiveness are major challenges, all the more so since increased private investments in R&D seem not to translate in productivity increases as expected.

Flanders is well known for its system of performance-based funding of university research (Debackere & Glänzel, 2004; Engels & Guns, 2018). At the occasion of the Nordic Workshop on Bibliometrics and Research Policy 2023, Engels & Guns analyzed the co-evolution of the PRFS with bibliometric performance indicators and reported an initial increase in per capita productivity. In the longer term, scholarly productivity and impact seem to have stabilized at a relatively high level, which also shows in the bibliometric indicators of the aforementioned regional innovation scoreboard. Holding such a position becomes less evident given the intense global competition for talent and infrastructure in science and technology, and may over time result in a slight or gradual reduction of competitiveness.

Less well known than the Flemish PRFS is that universities in Flanders have a legal obligation to assess the quality of their research activities (Engels et al., 2013). These research assessments resemble systems in the Netherlands and Norway (Sivertsen, 2017), whereby research assessment at the level of departments or research teams takes place without direct financial consequences for the university or the departments and teams involved. In other words, these assessments take place per

discipline and are intended as exercises to gather input and evidence on how to maintain and further improve quality and impact of university research. Like universities in Sweden (van den Besselaar & Sandström, 2020), universities in Flanders are autonomous in the organisation of these research assessments per discipline.

In addition, universities for many years also had a legal obligation to evaluate each university professor at least every five years (recently this legal obligation has been relaxed). Although Flemish universities were in principle free to decide on how to conduct such individual evaluations, some universities set up complex quantitative systems involving, among other elements, annual performance and goal setting reviews (DORA, 2023). The reform of those systems and the fact that all Flemish universities and the Flemish Rector's Conference (VLIR) where among the first signatories of the Coalition on Advancing Research Assessment (CoARA), illustrates that each of the Flemish universities is seeking a balance between expectations for productivity and impact, and nurturing academic freedom and diversity in research. In the next section, we delve deeper into how we balance these aspects in research assessments of teams at UAntwerp.

Research assessments of teams at the University of Antwerp

In 2007, the Research Board of the University of Antwerp decided to introduce a systematic external quality assessment of its research, through a discipline-specific approach and involving site visits by external peer reviewers (Engels et al, 2013). Since then, consecutive site visits took place according to a rotating system, in which each year the research groups belonging to two disciplines have been assessed. This way, all disciplines at the University of Antwerp have been evaluated twice since 2007. The Research Board opted for a protocol which is similar to the Dutch research assessment protocol (Standard Evaluation Protocol or SEP - since 2021 renamed Strategy Evaluation Protocol). Each international peer panel presents its assessments on four criteria – quality, productivity, societal engagement & impact, and viability - according to a five-point scale: (5) excellent, (4) very good, (3) good, (2) satisfactory, and (1) unsatisfactory. The panel provides a textual motivation for each score. Next to scoring the groups, the panel also reflects and provides feedback on the research policy of the department and/or faculty to which the groups belong and makes suggestions for the further development of research policy at the level of the department, the faculty, and the university (Houben, 2023).

The stated aim of the assessments is improving the quality and the impact of the research. By assessing the performance of the groups against their mission, strategy, and future plans, the panel members provide feedback on the past performance and current situation of each of the groups and are able to provide recommendations towards the future. Each assessment is to be regarded as a guiding principle, a means of self-reflection and positioning one's team in the research system. Although the units of assessment are the research teams, the assessment is strongly related to research policy within the department, the faculty, and the university. By assessing all groups within a department or faculty, the panel gets a broader picture and can make recommendations to each aggregation level where it sees fit. After all,

difficulties in the research agendas of the groups often are related to obstacles within the research policy on a higher level. As such, each assessment report provides each level recommendations by an international expert panel about the research itself, the research context and the research policy.
Ever since 2007 the assessment dossiers prepared in view of each visit emphasize a holistic approach, diversity, transparency, and validity. In the dossiers each research team provides qualitative context in the form of their mission, strategy, achievements, and a SWOT-analysis. Quantitative measures and indicators, that are known to and validated by the researchers in the discipline and each of the research groups prior to the submission of the preparatory documents to the expert panel, support these narratives. These quantitative indicators included information on inputs (e.g. overview of academic and technical staff, as well as doctoral and postdoctoral researchers; overview of funding acquired), process (e.g. duration of doctoral trajectories), as well as output (e.g. doctorates awarded; publications; patents), and impact (e.g. citations; spin-offs launched). In terms of scholarly outputs, the approach can be considered broadly in line with the recommendations of the Leiden Manifesto (Hicks et al., 2015), e.g. the inclusion of outputs in a diversity of languages and beyond international citation databases, and decided upon in consultation with the researchers in the discipline of focus. Still, the channels of publication and, where applicable, their impact factors are provided, as are personal bibliometric profiles of the professors in the group, leading to a possibility for focus on specific kinds of outputs (e.g. publications in high impact journals) over others. Over time we have put more emphasis on societal impact and incorporated information on Open Science practices, as well as on research integrity and a diversity of research outputs. The university research affairs office also applies a co-creation approach in the assessment process, by taking into account discipline specific characteristics and needs throughout the process. Such a way of evaluating has gained more and more attention since the creation of the SCOPE Framework (INORMS, 2021). UAntwerp professors also suggested potential panel members and chairs (Rahman et al., 2016). The entire process was also carried out in a transparent way by granting the professors and researchers access to all documentation and information with regard to the research assessment, including all the details behind numeric tables and graphs that the research affairs office prepares for the international expert panel (cf. Hong Kong Principles for assessing research, Moher et al., 2020).

Observations after two cycles of research assessments

In this section we provide a brief summary of observations after two cycles of research assessments at the University of Antwerp. We specifically zoom in on three aspects:

- The correlation of the assessment scores. As research groups receive scores on quality, productivity, impact and viability, we ask ourselves to what extent these ordinal scores correlate with each other. The higher the correlations, the more difficult panels might find it to differentiate these predefined dimensions during their assessments. Very high correlations may indicate a need to limit

the number of dimensions to assess, while dimensions that are less correlated indicate areas that might be in need of additional attention.

- The evolution of assessment scores from the first to the second cycle of assessments. Houben (2023) already reported, over halfway the second cycle, a clear increase of scores. Here we analyse this evolution upon the completion of the entire second cycle, and zoom in on the evolution of the scores for each of the four criteria.
- Lastly, we consider the main recurring issues that expert panels commented on, and what they might imply for the setup of the research assessments.

Correlation of scores

We calculate Spearman's rho correlations for the assessment scores within the first assessment cycle and within the second assessment cycle. Within each cycle, one expert panel per discipline assessed all the research teams within the given field. All correlations are positive and statistically significant ($p<.001$), yet the correlations in the second cycle are lower than in the first cycle, implying more variation of the scores per group in the second cycle. Especially in the first cycle, the high correlations seem to indicate the difficulties panels may experience to assess these predefined dimensions of the performance of teams separately. In the second cycle we still observe moderate correlations, although some panels were more inclined to e.g. assess impact and/or viability differently than the quality and productivity dimensions.

Table 1. Correlations of scores for quality, productivity, impact, and visibility in the first (upper right triangle) and the second (lower left triangle) cycle of research assessments.

Table	*Quality*	*Productivity*	*Impact*	*Viability*
Quality	-	.76**	.76**	.75**
Productivity	.60**	-	.73**	.65**
Impact	.43**	.49**	-	.73**
Visibility	.45**	.53**	.50**	-

Note. ** p <.001

Evolution of scores

We performed Mann-Whitney U tests for nonparametric assessment scores to evaluate the differences between scores awarded by the international expert panels to the research groups in the first cycle (N = 136) and those awarded in the second cycle (N = 112). The results indicate statistically significant differences between the scores on all criteria: quality ($z = 4.07$, $p < .001$), productivity ($z = 4.47$, $p = .001$), impact ($z = 2.34$, $p = 0.019$), and viability ($z = 2.27$, $p = 0.027$) with higher scores on these parameters in the second cycle (Figure 1).

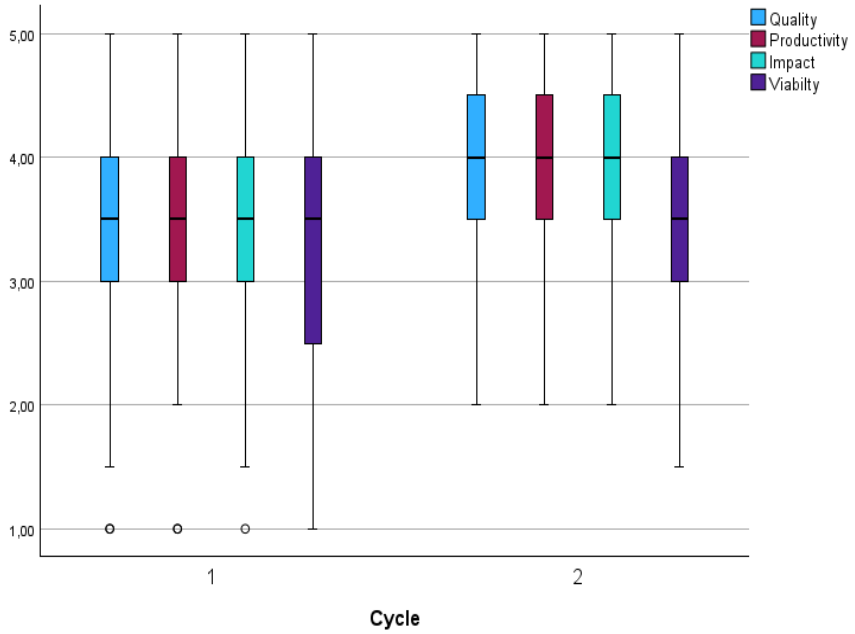


Figure 1. Scores per assessment criterion during the 1st (136 groups) and the 2nd cycle (112 groups) of research assessments.

Van Drooge et al. (2013) observed a similar inflation of scores in the Netherlands. Since the members of an expert panel in the second cycle receive the assessment report from the first cycle, they often make comparisons with previous recommendations and scores. As many groups try to live up to these recommendations, panel members tend to reward the groups with a higher score than in the previous cycle. Indeed, the intensity of research at Flemish universities, e.g. in terms of number of researchers per professor, has steadily increased at least until 2020, thus providing support for higher scores in terms of productivity. At the same time, the evidence is mixed at best when it comes to per capita productivity of research outputs, or quality and impact of research. Therefore we consider the gradual increase of scores for quality, productivity, and impact likely to represent a learning or habituation effect, whereby the teams learn to position themselves more strategically whereas the assessors reward positioning and results that are in line with their expectations and recommendations.

The increasing scores for viability, however, probably also represents another effect, that of further clustering of research groups. Indeed, even though the UAntwerp added two new faculties (in Applied Engineering and in Design Sciences), and two new departments (in Revalidation Sciences and in Applied Linguistics) in 2014, the number of research groups in the second cycle is considerably lower than in the first cycle, mainly due to mergers of groups into somewhat bigger wholes. Indeed, with the current total of 112 research groups, the average group now brings together 4 full time equivalent of professorial appointments, while this used to be 3 FTE.

Recuring themes in the self-assessments and the expert panel reports

In terms of recurring themes addressed in the expert panel reports, we distilled five main themes across all assessments. A major observation is that very few aspects of these themes seem to relate to one of the assessment dimensions specifically.

Not surprisingly, a first recurring theme concerns the attraction and retention of talent. Indeed, research cannot do without talented and well-trained researchers at all levels, from PhD candidates to professors. Hence research groups and departments often brought up this challenge, while the expert panels repeatedly stressed the importance of investing in early career researchers, their training and their career paths.

Secondly, research groups and departments frequently brought up funding for research as a theme, often with concerns regarding the high competition for grants and fellowships, and the extensive uncertainty that comes with low success rates at all levels (e.g. at the Flemish and European level). We find this also in the panel reports, in particular the recommendation to strengthen support in the pre-award phase.

Thirdly, the units frequently commented on the need for state of the art research infrastructure, including the cost of maintenance of such infrastructures. The expert panels for their part underscored the importance of research infrastructure, while also emphasizing the need for collaboration and efficiency gains, and the need to prioritize long-term investment in technical staff and infrastructures.

Fourthly, the high workload of researchers and professors is a recuring theme, attributed to a range of issues such as administrative overload and teaching load. Indeed, the expert panels recommended several times to (re)balance teaching and research, while also suggestion increasing administrative support.

We note that the expert panels tend to link these themes to the future research performance of the teams, although rarely specifically to the dimensions of quality, productivity or impact. Just like the medium to high correlations of these assessments, this seems to illustrate the impossibility for the expert panels to disentangle these different dimensions during the assessments. The same holds for other recuring recommendations, such as the suggestion to strengthen the international profile and networking of the university.

The expert panels linked only a few themes or topics explicitly to quality, productivity or impact of the research. For example, some panels recommended to aim explicitly for high impact journals, while others encouraged to focus more on societal impact. An often recurring theme was the need to facilitate and stimulate interdisciplinary research and to tackle the hurdles that researchers experience to engage in such research, which panels often linked to the innovativeness and (future) societal impact of the research.

Overall we observe that the main recurring themes in the expert panel reports, across all disciplines and across the two assessment cycles, rarely relate directly to one of the dimensions of the assessment. Rather they tend to relate to the research context, the organisation, and the (research) policies at the departmental, the faculty, and the university level.

Towards a new framework of responsible research assessments

Currently we are preparing, in close interaction with the Research Board of the university, the third cycle of research assessments. In line with the SCOPE framework (INORMS, 2021), a first focus of these discussions concerns the purpose of the research assessments. In particular, we suggest a refined notion of the purpose of the research assessments. Instead of the purpose 'to improve the quality and the impact of the research', we suggest a more specific purpose 'to contribute to a context that is conducive to high quality research with high impact on science and/or society'. This explicitly includes recognizing values of academic freedom, diversity and inclusivity of research, and contribution to the prosperity and well-being in the region and beyond.

In order to attain this ambitious goal, we intend to prioritize a thorough understanding of the context of the research as a starting point for all assessments. This will include both the broader context of research in Flanders, as well as continuous monitoring through dashboards of each research unit's performance in terms of inputs, process, outputs, and impact. We aim for more ethical and responsible use of all available indicators, by positioning them explicitly as background information to the mission and strategy of each research team and making this information permanently available to each team. This context per research team will provide the background for an analysis of strengths, weaknesses, threats, and opportunities by each cluster that is to be assessed. This qualitative SWOT analysis will then serve as the main input to a panel of experts.

In terms of aggregation level, we intend to evolve towards broader clusters, larger then one discipline or department and perhaps sometimes involving several faculties at once. Hence the research teams will become the building blocks of an assessment, rather then the units of assessment. What will be assessed, with a view of maximal alignment, is the research context, the organization, and the research policy at the level of the departments, the faculties, and the university. This 'assessment' will not be a numerical assessment, but will take the format of a set of recommendations, in order to foster collaboration, and aligned strategies at all levels, taking into account the needs of a variety of stakeholders, from early career researchers, to professors and research group leaders as well as heads of department, deans, and the rector team.

Conclusions

The University of Antwerp conducts research assessments of research teams since 2007. Over the years the approach of the assessments has evolved, e.g. gradually paying more and more attention to context and process elements. With the third cycle of assessment in preparation, we advocate a thorough rethink of the assessment approach, refocusing on the research context, the organization, and the research policy at the level of departments, faculties, and the university in order to align more with the ambitions of responsible research assessments. At the occasion of the ISSI 2025 Special Track FRAME, we will reflect on our experience of conducting research assessments since 2007 and the state of play of the preparation of a third cycle of responsible research assessments. The focus of our presentation will be the

challenge of integrating contextual quantitative indicators and qualitative SWOT analyses in the assessment of research.

Acknowledgments

We wish to acknowledge all researchers and local and international professors that have contributed to the research assessments at the University of Antwerp. Furthermore we wish to acknowledge the many colleagues in Flanders and beyond that have provided us input and reflections on how to conceptualize and organize responsible research assessments.

References

Abramo, G. (2024). The forced battle between peer-review and scientometric research assessment: Why the CoARA initiative is unsound. *Research Evaluation*, 1–8. https://doi.org/10.1093/reseval/RVAE021

CoARA. (2022). *Coalition for Advancing Research Assessment*. https://coara.eu/agreement/the-agreement-full-text/

Debackere, K., & Glänzel, W. (2004). Using a bibliometric approach to support research policy making: The case of the Flemish BOF-key. *Scientometrics*, *59*(2), 253–276. https://doi.org/10.1023/B:SCIE.0000018532.70146.02

DORA. (2023). *Case study: Ghent University*. https://sfdora.org/case-study/ghent-university/

Engels, T. C. E., Goos, P., Dexters, N., & Spruyt, E. H. J. (2013). Group size, h-index, and efficiency in publishing in top journals explain expert panel assessments of research group quality and productivity. *Research Evaluation*, *22*(4), 224–236. https://doi.org/10.1093/reseval/rvt013

Engels, T. C. E., & Guns, R. (2018). The Flemish Performance-based Research Funding System: A Unique Variant of the Norwegian Model. *Journal of Data and Information Science*, *3*(4), 45–60. https://doi.org/10.2478/jdis-2018-0020

Global Research Council. (2024). *Dimensions of Responsible Research Assessment*. https://globalresearchcouncil.org/about/responsible-research-assessment-working-group/dimensions-of-rra/

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. https://doi.org/10.1038/520429a

Houben, B. (2023). 15 years of research assessment exercises at the University of Antwerp: How evolution in the research landscape leads to change in the assessment practice. *STI 2023 Conference*.

IDEA Consult. (2024). *Systeemanalyse Onderzoek, Ontwikkeling en Innovatie en uitgaventoetsing van het 'Vlaams beleid in het kader van productiviteit'*. Departement Economie, Wetenschap & Innovatie.

INORMS. (2021). *The SCOPE framework. A five-stage process for evaluating research responsibly*. Emerald Publishing. https://inorms.net/scope-framework-for-research-evaluation/

Moed, H. (2017). *Applied evaluative bibliometrics*. Springer International Publishing.

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., Coriat, A.-M., Foeger, N., & Dirnagl, U. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLOS Biology*, *18*(7), e3000737. https://doi.org/10.1371/journal.pbio.3000737

OECD Global Science Forum. (2025). *New expectations and demands from science: Rethinking research assessment and incentive structures*.

Rahman, A. I. M. J., Guns, R., Leydesdorff, L., & Engels, T. C. E. (2016). Measuring the match between evaluators and evaluees: Cognitive distances between panel members and research groups at the journal level. *Scientometrics*, *109*(3), 1639–1663. https://doi.org/10.1007/s11192-016-2132-x

Sivertsen, G. (2017). Unique, but still best practice? The Research Excellence Framework (REF) from an international perspective. *Palgrave Communications*, *3*(1), 17078. https://doi.org/10.1057/palcomms.2017.78

van den Besselaar, P., & Sandström, U. (2020). Bibliometrically disciplined peer review: On using indicators in research evaluation. *Scholarly Assessment Reports*, *2*(1). https://doi.org/10.29024/sar.16

van Drooge, L., de Jong, S., Faber, M., & Westerheijden, D. (2013). *Twintig jaar onderzoeksevaluatie: Feiten & cijfers* (p. 9). Rathenau Instituut. https://www.rathenau.nl/nl/werking-van-het-wetenschapssysteem/twintig-jaar-onderzoeksevaluatie

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. https://doi.org/10.13140/RG.2.1.4929.1363

Wouters, P., Glänzel, W., Gläser, J., & Rafols, I. (2013). The dilemmas of performance indicators of individual researchers-An urgent debate in bibliometrics. *ISSI Newsletter*, *9*(3), 48–53.

Zhang, L., & Sivertsen, G. (2020). The new research assessment reform in China and its implementation. *Scholarly Assessment Reports*, *2*(1). https://doi.org/10.29024/sar.15

https://doi.org/10.51408/issi2025_154

Toward Responsible Scientometrics: Normative Data Practices for Research Evaluation

Ying Huang[1M], Weishu Liu[2], Huizhen Fu[3], Jing Ma[4], Guijie Zhang[5], Yi Bu[6], Chao Min[7], Zhixiang Wu[8]

[1] *ying.huang@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan 430072 (China)

[2] *Weishuliu.@zufe.edu.cn*
School of Information Management and Artificial Intelligence, Zhejiang University of Finance and Economics, Hangzhou 310018 (China)

[3] *fuhuizhen@zju.edu.cn*
School of Public Affairs, Zhejiang University, Hangzhou 310012 (China)

[4] *majing@xidian.edu.cn*
School of Economics and Management, Xidian University, Xi'an 710126 (China)

[5] *zgjzxmtx@163.com*
School of Management Science and Engineering, Shandong University of Finance and Economics, Jinan 250014 (China)

[6] *buyi@pku.edu.cn*
Department of Information Management, Peking University, Beijing 100084 (China)

[7] *mc@nju.edu.cn*
School of Information Management, Nanjing University, Nanjing 210023 (China)

[8] *cnwzx2012@njtech.edu.cn*
School of Economics and Management, Nanjing Tech University, Nanjing 211800 (China)

Abstract

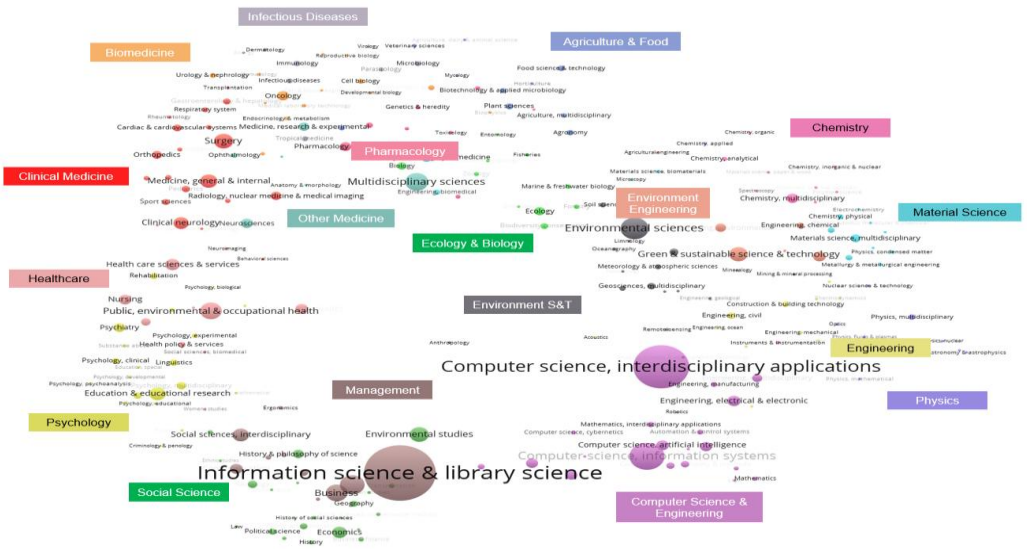
Conducting scientometric research in a rigorously normative manner requires careful attention to responsible metrics and data practices throughout the research life cycle. This study offers practical suggestions from a data perspective, addressing the key stages from source selection to data sharing while emphasizing transparency and policy implications for evaluation systems. First, the selection of data sources necessitates the alignment of data quality, inclusion criteria, and research objectives. Researchers must ensure that dataset characteristics align with their analytical goals and accurately document sources, including aspects such as coverage and granularity. Data retrieval and acquisition require a clear understanding of the database's scope and update mechanisms. A deliberate approach—utilizing tailored search terms, following platform-specific syntax, and transparently reporting retrieval strategies—ensures reproducibility. Data pre-processing involves a clear comprehension of the field and specific tasks (such as deduplication and normalization). Standardized protocols should connect raw data to analytical requirements while maintaining transparency in methodological choices. For analysis, selecting appropriate tools and differentiating standardization methods (e.g., citation versus collaboration network metrics) is critical to support responsible metrics in evaluation systems. Visualization should enhance interpretability without distorting evidence. Interpretation must strike a balance between robustness and restraint, avoiding over-claiming patterns or neglecting contextual nuances. Engaging domain experts helps mitigate disciplinary biases. Data

storage requires the systematic documentation of fields, derived variables, and backup protocols to ensure auditability—a key aspect of transparent and accountable research practices. Finally, obligations for sharing and citations include disseminating data through repositories and formally crediting open-access datasets to maintain scholarly standards. By integrating these guidelines, researchers can enhance scientometrics' methodological rigor, replicability, and ethical accountability, contributing to more responsible and policy-relevant evaluation systems. This framework highlights how normative advancements depend on meticulous data stewardship at each procedural stage, with broader policy implications for research assessment.

Introduction

Scientometrics is an interdisciplinary discipline that uses quantitative methods to quantify all aspects of science and, as a whole, to reveal its development patterns. It is an important branch of the Science of Science domain and an active field in current scientific research. The term "science" here refers not only to science as a body of knowledge but also science as a social activity, establishment, and industry. Scientometrics, as used in this study, is a broad concept that emphasizes the use of quantitative methods to explore the laws of scientific development and, thus, it also includes informetrics and nascent altmetrics.

Scientometrics reveals the current state of development of a disciplinary field or research topic. As such, it can help researchers and research managers quickly understand a field's full picture. Simultaneously, increasingly sophisticated methodologies and software tools have further expanded the boundaries and groups of users of scientometrics, making it possible for scientometrics-related research to be applied beyond existing disciplines in a variety of fields, such as environmental science, psychology, and clinical medicine (as shown in Figure 1).



Note: TS=(Scientometrics OR Bibliometrics OR Informetrics) AND DT=("REVIEW" OR "ARTICLE") AND FPY=(1978-2024)

Figure 1. Distribution of journals involved in research papers related to scientometrics.

However, as scientometrics has expanded and developed, problems such as ambiguous data source selection, generalized data retrieval, lack of data pre-processing, misleading data analysis and visualization, arbitrary data interpretation, confusing data storage, and unclear representation of data sharing and citations have remained prevalent. To this end, this study aims to provide suggestions from a data perspective that help ensure that scientometrics research is stable, diverse, transparent (Bornmann et al., 2021), and reproducible (Velden et al., 2018; Waltman et al., 2018). Moreover, these practical recommendations are closely aligned with the broader normative debates on responsible metrics in research evaluation. By integrating insights from recent frameworks such as the Coalition for Advancing Research Assessment (CoARA), this study underscores the importance of adopting rigorous data practices to support evaluation reform efforts. Such alignment not only enhances the relevance of this work to the objectives of the Framework for Responsible Assessment Metrics in Research (FRAME) but also contributes to fostering a more equitable and sustainable research evaluation ecosystem.

Addressing Normative Challenges in Scientometrics Through Data Practices

Data source selection

The selection of data sources has a fundamental effect on scientometric research. Improper data source selection significantly affects the reliability of the conclusions drawn. In recent years, both the new bibliographic databases (e.g., Dimensions, Semantic Scholar, Crossref, OpenAlex, and OpenCitations) (Gusenbauer, 2022) and the new variables provided by traditional databases (such as "funding acknowledgment", "usage count", "enriched cited references" in the Web of Science) have enriched the user's range of data source selection. It is worth mentioning that open data sources help make scientometric research more transparent and reproducible, and they also help make the field of scientometrics more equitable, since they enable scientometric research to be carried out by researchers and institutions that cannot afford a subscription to commercial proprietary data sources. However, although rich data sources provide opportunities for scientometric research, they also challenge scientometric scholars in selecting appropriate data sources. Therefore, the following recommendations were made.

Data quality must meet the research's needs

The increasing diversification and greater availability of data sources have given scientometricans more choices, but the quality of these new data sources may not always meet our expectations. Even when the most widely used commercial databases are chosen as data sources, such as Scopus or the Web of Science Core Collection (Zhu & Liu, 2020), the studies have shown different types of errors. Some scholars have even described Scopus as a "museum of errors/horrors" (Franceschini et al., 2016). For example, Scopus has a substantial number of funding information errors, and users need to carefully check the data accuracy before using the Scopus database for funding analysis (Liu et al., 2020). With all data sources, such as Google Scholar, users first need to gauge the quality of the database in terms of accuracy,

completeness, or whatever the appropriate metric and only reasonably use that data source on the premise that its quality meets the needs of the research.

Inclusion criteria must meet the research's needs

The content inclusion criteria of the selected bibliographic database(s) should also meet the research needs. Some bibliographic databases are selective (e.g., Web of Science), whereas others are not (e.g., Google Scholar). There are also differences in the selection criteria for including data sources. For example, the Emerging Sources Citation Index (ESCI), which has a large number of journals, is a collection of potential candidates for the other three classic journal citation indices of the Web of Science Core Collection, but its selection criteria are slightly lower than those of the other three authoritative indices (Huang et al., 2017). Hence, users should judge whether the criteria for including some literature in a specific data source meet the requirements of the research. Sometimes, a tradeoff between quality and quantity may need to be made.

Data characteristics should match the research question

A mismatch between the characteristics of the data source and the research question may lead to anomalous research conclusions that further mislead the practice. Several types of bibliographic databases are frequently used. These include multidisciplinary databases (e.g., Web of Science, Scopus, Dimensions, Crossref, and OpenAlex), disciplinary databases (e.g., Medline for medicine), and regional databases (e.g., the Chinese Science Citation Database, the Chinese Social Sciences Citation Index, the Russian Science Citation Index, and the SciELO Citation Index (South America). Different data sources also have language, regional, and discipline biases (Liu, 2017; Martín-Martín et al., 2021). Users need to be familiar with the characteristics and shortcomings of the data sources they consider, and choose the appropriate data sources according to an accurate analysis of the needs of the research problem. For example, if you plan to study the journal paper output of social science research in China, in addition to the Social Sciences Citation Index, the Chinese Social Sciences Citation Index is an important data source. Another example is that it is not appropriate to use acknowledgments from the Web of Science Core Collection to analyze funding in the social sciences prior to 2015. Web of Science began to include funding information for SCIE papers in 2008, but funding information for SSCI papers has only been systematically recorded since 2015 (Liu et al., 2020).

Describe the data source precisely

An accurate description of the data sources is the cornerstone of research reproducibility. However, some studies have found that even in the field of scientometrics, there are still serious problems with the descriptions of the data sources used in many studies (Liu, 2019). One prominent problem is the failure to accurately and unambiguously represent the Web of Science and its core collection. Web of Science is now a search platform provided by Clarivate, which contains databases such as the Web of Science Core Collection. However, the full Web of

Science Core Collection contains two chemical indices and eight citation indices (one of which is the Science Citation Index Expanded). Different institutions may subscribe to different subsets of core collections with different years of coverage. Therefore, when using the Web of Science Core Collection, the sub-datasets used and the corresponding coverage year information need to be clearly disclosed (Liu, 2019).

Data retrieval and acquisition

As the basis of subsequent data analyses, accurate and proper data retrieval and acquisition directly influence the credibility and effectiveness of the data analysis and the corresponding results. Improper data retrieval and acquisition strategies might not only lead to misleading and distorted information but also waste human, energy, and material resources. Appropriate data retrieval and acquisition strategies are built on adequate investigation and a solid understanding of both databases and research needs. This means being familiar with the advantages and limitations of databases and their search rules. We investigated the common problems with data retrieval and acquisition in existing scientometric studies and summarized the issues into the following four precautions and norms for researchers to pay attention to when engaging in scientometric studies.

Be familiar with the scope and update rules of the database

The different backgrounds and service objects of the different databases indicate that each database has its own preferences when it comes to which data to collect. Therefore, a good retrieval strategy should not exceed the scope of the database. Taking the Web of Science as an example, the SCI database began collecting abstracts, author keywords, and keyword-plus information in 1991 (Clarivate Analytics, 2020). Therefore, when the topic filter (topic includes title, abstract, author keywords, and keyword-plus) is used in SCI databases, it is recommended to avoid the watershed year of 1991 during the investigation period. Otherwise, there might have been a misleading jump in the number of publications in 1991, skewing growth trends (Liu, 2021).

In addition, it usually takes time for a database to collect the newest data, and most databases have a fixed update frequency. For example, Web of Science generally updates on a bi-weekly basis. In addition, most databases have a time lag. Hence, the search date can also influence search results. For example, if we want to find documents published in 2020, a search on 1 January 1, 2021, will find fewer documents than a search on June 30, 2021, even if the same search strategy was used. Generally, a good buffer zone for accommodating the time lag is three to six months after the end of the study period.

Deliberately select search terms

Terms in a topic search are a common way to retrieve publications in a target field. Appropriate search terms can help ensure the accuracy of the data retrieval. A preliminary group of search terms should be selected only after the target field and the needs of a relevant search are fully understood. This preliminary search strategy

can be further improved through discussions with relevant experts and multiple retrievals with spot checks and tests of the search results (Arora et al., 2013).

Evaluating a search strategy often requires an appropriate balance between recall and precision. Generally, when precision is high, recall is often low, and vice versa (Lambert, 1991). Researchers must also pay attention to the abbreviations used in the search terms. For example, the abbreviation of artificial intelligence is abbreviated as AI, but if AI is used as a search term, literature related to the biological index AI and the medical index AI would probably also be found (Wilson et al., 2022). In fact, it is very difficult to ensure that abbreviations are used only in the expected field. Generally, documents with the same abbreviations in nontarget fields are also found. Therefore, if precision is given priority, abbreviations need to be selected cautiously, and if abbreviations cannot be avoided, checking and cleaning the search results will be necessary to exclude irrelevant documents.

Attention should also be paid to the hypernyms and hyponyms of the search terms, and choosing a group of appositives is also recommended in one's search strategy. For example, (Fu & Waltman, 2022) used several appositives of climate change, including "climate change *", "climate change *", "climate variability *", "climate variability *", "global warming", "climate warming", "climate warming," in their data retrieval strategy to assemble a comprehensive body of literature on climate change. Alternatively, if the hyponym "greenhouse" of climate change research were to be included in the search terms, the search may result in a greater proportion of documents related to greenhouse effects, and this could overweight the characteristics of greenhouse across the whole field, skewing the research results.

Understand the search rules of the database

A data retrieval strategy and a detailed search query should be constructed based on a specific database. Before conducting a search, it is important to be familiar with the rules of the selected data platform. Common search rules include Boolean operators, proximity operators, quotation marks, wildcards, and truncation. The negligence and misuse of search rules have a significant impact on data retrieval and further analysis. Taking the use of quotation marks as an example, Topic=(solid waste *) and Topic=("solid waste *") are two different search queries. The first search query found documents with solids and waste separately in different sentences. Thus, some documents irrelevant to solid waste were likely to be found. However, the second search query with quotation marks only found papers with solid waste as a phrase. This mistake was made in (Liu et al., 2014) and, as a result, almost twice the number of publications was found as there should have been. There were also significant differences in subsequent results, such as growth trends and country performance. In this case, the lack of quotation marks leads to inaccurate results and unreliable conclusions.

Elaborate the strategy of data retrieval and acquisition

The detailed strategies for data retrieval and acquisition need to be explained in the final publicly released publications to ensure the transparency of data retrieval and acquisition. A clearly stated search strategy provides the basis for judging the

reliability of scientific activities. It also helps ensure the authenticity and certainty of scientific knowledge. It is suggested that the general description of the search strategy specifies the database, subdatabase, search query, investigation period, actual retrieval time, and any other salient information. The basic criterion for judging whether a search strategy is clearly explained is that if a reader were to follow the explanation, they would retrieve approximately the same dataset and reproduce similar search results. Even though data sources such as Web of Science, Scopus, and others are continuously being updated, new records are being added, and old records may be updated or may even be removed by documenting the search strategy, other researchers may be able to approximately reproduce the analysis.

Data pre-processing

Data pre-processing is defined as a series of necessary processes that occur after data retrieval and before data analysis and modeling. It primarily includes tasks such as data integration, data cleaning, and data transformation. For instance, multi-source data are often used in scientometric research, but the structure of the data gathered from each source can be inconsistent. Simultaneously, with the increasing amount of scientific data, problems such as missing data, duplicate data, and anomalies in the data may seriously affect the validity and reliability of data analysis results. High-quality data pre-processing can effectively improve the quality of the analysis results and conclusions. It can also make the data fit the models and analysis tools better, which can help improve the efficiency of the analysis. Therefore, we have provided the following suggestions for ensuring efficient and high-quality data pre-processing.

Adequately understand the data fields

Owing to the inconsistency of field formats and statistical calibers, data fields from different data sources must be merged carefully. For example, the "Times Cited" fields of different databases are usually limited to the internal statistics of their respective databases, which means that the counts given by different databases are likely to be different. Therefore, the integration of such fields needs to be handled with caution (Pech & Delgado, 2020). Therefore, the characteristics of a field should be fully considered during pre-processing. For example, there are several fields regarding funding information in the Web of Science, among which the "FU" field records the funding institution and authorization number, while the "FX" field records funding information as unstructured text. Hence, the "FU" field is better structured and, therefore, less difficult to pre-process, whereas the "FX" field provides more complete information. In addition, keywords, as subject tags of research content, are also commonly used in scientometrics research. However, there are two keyword fields in WOS, "Author's Keywords" and "Keywords Plus". Unlike the Author's Keywords, the Keywords Plus field is based on the original records and their references, and is designed to make data retrieval more convenient. Therefore, in some research fields, Keyword Plus may not be sufficient to describe the uniqueness and comprehensiveness of the content of the literature and should be used with caution (Zhang et al., 2016).

Clarify data pre-processing tasks

In scientometric research, the original field content retrieved from a database cannot usually be directly used for subsequent analyses. Rather, the data needs to be pre-processed. Pre-processing tasks generally include data integration, deduplication, field extraction, and cleaning. However, in practice, the pre-processing methods to be applied should be selected according to the requirements of the specific analysis. Data deduplication and merging are basic pre-processing operations. Even when data have only been collected from a single database, there can still be problems with duplicate records. For example, the same paper may be included with different statuses, such as published, online, in-press, correction, and withdrawal. Duplicate records should be merged and deleted (Gagolewski, 2011).

Another common pre-processing task is to disambiguate author names, institutions, and references. If the research includes author analyses, author disambiguation should be considered first. This includes distinguishing author names and merging the different record forms of the same author. Different disambiguation strategies may impact the subsequent cooperative network construction or citation analysis (Kim & Diesner, 2015, 2016). Disambiguating Asian authors, especially Chinese and Korean authors, is generally considered more challenging because of regional and linguistic differences (Harzing, 2015; Strotmann & Zhao, 2012). Therefore, the author's other attributes, such as ORCID, institution, email address, research field, and ResearchGate records, can be considered to further improve disambiguated results (Abdulhayoglu & Thijs, 2017; Han et al., 2017; Porter, 2022; Youtie et al., 2017)

Ensure the standardization and repeatability of data pre-processing

Data pre-processing often relies on rules, stop word lists, and even manual processing to repeatedly revise and improve the results. Standardizing and making the data pre-processing procedure repeatable are key to ensuring the robustness and reliability of data analysis.

Taking text feature extraction as an example, common pre-processing steps include stemming, stop word list filtering, TF-IDF, fuzzy matching, and others. However, because text analysis usually requires consideration of the characteristics of disciplines or domains, different cleaning methods or processing steps may affect which text features are extracted (Newman et al., 2014; Zhang et al., 2014). For example, the word "nanotechnology" may be highly valuable in multidisciplinary studies. However, it is probably meaningless in a study of nanotechnology and may even be classified as a stop word. Therefore, data pre-processing should ensure that the procedural steps are rational in terms of stop word list design and threshold selection. Moreover, although in practice, it is usually not possible to document data-processing procedures in full detail in a research paper, making source codes or documentation openly available seems to be the best solution. It should also be emphasized that data pre-processing is an almost endless task of improvement, but it can be very time-consuming. Therefore, it is usually necessary to balance the quality of pre-processing with time, labor, and cost taken to do it. As the volume of

pre-processing tasks increases, the results can be evaluated through sampling or other metrics to ascertain the quality and efficiency of the pre-processing regime.

Bridge data pre-processing results to data analysis requirements

The typical results of data pre-processing are cleaned data, extracted relationships, and preliminary statistics, which provide fundamental information for subsequent analysis tasks. For example, social network analysis tools, such as Gephi and Ucinet, have different requirements regarding the structure of the network data to be input, which can be either in the form of an edge or an adjacent matrix. Therefore, the structure and storage format of the pre-processing results should be considered to ensure that they fit the subsequent analysis tools and analysis methods. This will also help improve the efficiency of the data analysis.

Data analysis

Data analysis is often carried out by means of some data analysis methods, tools, or software. Whichever means, the analysis should be rigorous and standardized, and the process should be as clear and transparent as possible. Anything less may lead to misjudgments in the subsequent data interpretation and could also raise the threshold for reproducibility of the research. Based on the common problems with data analysis in existing scientometrics research, we suggest that more attention should be paid to the selection of data analysis software, the use and parameter settings of any software used, data standardization, and visualization of data analysis.

Select appropriate data analysis software

Currently, many analytical tools can be used in scientometric research, so users must choose the appropriate one according to the characteristics of their data and the purpose of their analysis. For example, CiteSpace and VOSviewer can be used for co-citation, co-authorship, and keyword co-occurrence analysis. CitNetExplorer and HistCite were used for direct citation networks. Pajek, Ucinet, and Gephi can be used to visualize social network data and analyze network characteristics. Each tool has its advantages and limitations, and no single piece of software can accomplish all functions. Therefore, multiple tools must be used simultaneously. In addition, existing scientometric software has its limitations, and sometimes, researchers need to rely on general-purpose tools, such as Python and R. However, when using software, record which version of a software tool was used and which parameters were adjusted, which is very important for the optimization of the network, and explain the meanings of the parameters. For example, the layout area of VOSviewer provides "Attraction", "Repulsion", and "Advanced Parameters" to optimize the layout of a graph. Different parameters produce different rendering effects.

Distinguish different data standardization

In scientometric research, many similar issues related to data standardization require attention, and the counting method is one example of a broader set of issues. This refers to the calculation method of assigning ownership of scientific research papers according to certain rules. The counting method can be used to calculate the number

of papers published by authors, institutions, and countries and to analyze the frequency of citations. Most notably, this method can influence the allocation of scientific research resources and the content of science and technology policies (Sivertsen et al., 2019).

There are many ways to categorize these counting methods. The basic methods used are full counting and fractional counting. However, based on the correspondence between counting objects and counting units in various metrological problems, Gauffriau et al. (2007) divided the various counting methods into five categories: complete counting methods, complete-normalized counting methods, straight counting methods, whole counting methods, and whole-normalized counting methods. These can also be divided into different counting units. Here, we have full counting, country/organization/address/author level fractional counting, first author counting, and corresponding author counting (Waltman & van Eck, 2015). Full-counting methods reflect participation, whereas fractional-counting methods reflect contribution. Full counting is more commonly used at the individual level, as in the h-index proposed by Hirsch (Hirsch, 2005). Fractional counting is an aggregate metric in which the sum of the counts is the same as the number of papers. As such, this style of counting provides balance, consistency, and accuracy in standardized bibliometric measurements. In scientific research evaluations, it is usually necessary to compare the academic influence of papers in different fields. However, because different fields have different citation densities, these counts must be standardized before undertaking such a comparison. Moreover, the counting method used affects the results of standardization across different disciplines. Using different counting methods to calculate the number of citations in co-authored papers will result in different results, which will further affect any standardized citation impact indicators calculated when comparing the influence of papers across different fields (Lin et al., 2013).

Give full attention to data visualization

To make good use of graph functions in analysis tools, appropriate dimensions and quantities must be selected. Only this way will one create a picture that is "worth a thousand words." On this basis, the selected visual map should be based on how best to present the data. For example, Citespace provides a network view, timeline view, and time-zone view. VOSviewer offers a network view and cluster density view. Pajek has 2D, 3D, and dynamic community maps, whereas SciMAT provides an item overlay map, an evolution map, and a cluster network.

In addition, there are other issues that need to be addressed when performing data visualization. Different authors often use different words for the same concept, so it may be necessary to combine synonyms such as e-health, eHealth, mobile learning, and m-learning; otherwise, the graph obtained will be neither refined nor accurate. There are often similar problems with the names of countries and institutions, such as England, the UK, the University of Tokyo, and Tokyo University. In the case of VOSviewer, the above problem can be normalized based on the vocabulary obtained before visualization.

Data Interpretation

Reasonable and accurate interpretations are mandatory, regardless of the data analysis methods, tools, and/or software applications. Interpreting data requires a scientific nature and constructiveness. Here, a scientific nature means that one should strictly and standardly follow certain rules, regulations, and/or procedures of data interpretation and understand that inappropriate interpretations can greatly distort the results of data analyses. Constructiveness, on the other hand, is reflected by the fact that people from different domains may render various results because of factors such as professional background and subjective feelings. Hence, we suggest the following rules when interpreting data:

Preventing over-interpretation

There are many examples of overinterpreting data in scientometric research, among which people often confuse the difference between correlation and causation (Pearl & Mackenzie, 2018). At present, the majority of scientometric studies have been undertaken at the correlation level. This includes, for instance, (1) presenting the correlations found between two variables with some statistical charts (e.g., bivariate scatter plots, tri-variate bubble plots, etc.) and/or (2) depicting the results of statistical regression analyses between multiple variables. In studies indicating correlation-level results, one should avoid using words that imply causation when interpreting the results. For example, words such as "result in", "lead to", "influence", "impact", and "affect" all hint at causation. If causal inference has not been tested in the research, more careful and conservative expressions should be used. These include, but are not limited to, "relate to", "is proportional to", "as X increases, Y tends to...", etc.

In addition to confusing correlation and causation, scientists conducting descriptive research (e.g., mapping knowledge domains) may have made some inherent assumptions. These preconceived notions might lead them to focus on one part of the results that matches their psychological expectations when interpreting the results and ignore those that do not. Avoiding this pitfall requires vigilance when interpreting data.

Avoiding under-interpretation

At the other extreme, over-interpreting data is under-interpreting data. Generally, interpreting the results of scientometric research can be divided into four levels:

Level 1: Numerical level. Here is an example of an interpretation using the numerical dimension: "In biology, when the proportion of publications' references not belonging to the discipline is 0, the normalized value of their citations is about 0.5". This level is the most straightforward and serves as the basis for further interpretation. It should be noted that when interpreting at the numerical level, it is also necessary to focus on outliers, trends (temporally), multivariate relationships, and so on.

Level 2: Numerical comparison and inductive level. A typical example of Level 2 could be something like "as the value of the horizontal axis increases, the value of

the vertical axis increases first and then decreases." Note that Level 2 still does not touch on any conceptual level of scientometric knowledge.

Level 3: Conceptual level. This layer links the numerical results (indicators) to scientometric concepts and/or domain knowledge. Suppose that the indicator on the horizontal axis is the proportion of a publication's references that do not belong to a discipline. The corresponding concept/construct of this indicator could be the degree of interdisciplinarity of a publication. If the indicator on the vertical axis is the relative number of citations of a publication in its discipline, the corresponding concept/construct of this indicator could be a publication's scientific impact. Considering the Level-2 interpretation we mentioned, we may further interpret it as, "As the interdisciplinary nature increases, the academic influence of scientific literature first increases and then decreases."

Level 4: Implications. Discussions at this level mainly focus on the significance and importance of the findings from Levels 1-3. Level 4 interpretations often vary and can be carried out from many distinct aspects, such as theories of scientometrics, methodology, scientific and technical policies, and pedagogy. For example, " with the increase of interdisciplinarity, the scientific impact of publications first increases and then decreases." may offer empirical evidence if it presents a causal relationship to help policymakers and funding providers form policies on interdisciplinarity and unidisciplinarity.

Understanding these four levels of data interpretation would help researchers paint a more vivid and comprehensive picture of the material under study. In addition, preventing underinterpretation generally requires a comparison with previous research results.

Involving domain experts in the interpretation

Scientometrics is a typical "meta-discipline". This means that empirical research in scientometrics often involves one or more disciplines. Therefore, domain experts are often required to interpret data. If one traces back the history of citation analysis and mapping knowledge in domain studies (e.g., co-citation and bibliographic coupling analyses), for example, one impressive highlight worthy of recognition is that domain experts have strictly interpreted the results rather than simply discussing results without the perspective of any domain knowledge. Otherwise, normativeness, rigor, and persuasiveness can also be significantly reduced. More severely, the lack of interpretation by domain experts may hinder further reflection, optimization, and innovation of research methods (Bar-Ilan, 2008).

Data storage

The organization and storage of data are essential steps in a bibliometric analysis. This kind of housekeeping ensures that one obtains reliable quantitative analysis results and that the results are repeatable (Ferrara & Salini, 2012). Data storage runs through the entire scientometric research process and plays an essential role in academic research and communication, manuscript submission and revision, and even after publication. Therefore, we make the following recommendations when storing data.

Record all kinds of data fields and their derived variables

There is a range of metadata fields for scientific and technological documents, each with rich meaning (Pranckutė, 2021). However, for convenience, these fields are often labeled in the form of abbreviations, making it difficult to determine their meanings from the name. For example, in the core collection field of Web of Science, there are at least four fields concerning citations and usage: TC indicates the number of citations in the core collection, Z9 indicates the citations in the core collection, U1 is the number of times used in the last 180 days, and U2 is the number of times used since 2013. Another example is the literature on patent type. The applicant and inventor have relationships with the subjects related to the patent right, but the former indicates the patent applicant (both the institutions and the individuals), and the latter indicates the patent's inventor (generally a specific person). These two are easily confused. The above are only original data fields derived from the scientific literature. In some applied follow-up research, new variables based on the original data fields will be generated, and some new variables will result in new versions owing to differences in algorithms and parameter settings. In this way, managing the data fields and derived variables may become more complicated, and errors may occur if one is not careful. Furthermore, as time goes by, it is easy to forget the specific meaning of fields and variables or become confused about them. Therefore, it is wise to provide each metadata field with a clear name and definition to avoid this problem. Moreover, if the variables for subsequent analysis have been generated based on metadata fields, it is even more necessary to name the fields and variables scientifically to ensure that they are scalable and that new variables can be generated and added.

Deal with data backups and file preservation

Data from scientific literature have increasingly become a support for research results and the basis of scientific measurements. Therefore, long-term preservation and archiving of data is becoming a realistic demand for scientometric analysis. However, this raises many logical questions. Who will maintain a regular data backup? Where are data archives stored? If funding agencies are present, what specific requirements do they have regarding the format and scheme of data retention? All these questions must be considered. If only a small amount of data is available, a local storage solution may be sufficient. However, if there is a large amount of data, it may be better to choose a cloud platform as the storage medium.

Data security is also worth noting (Assunção et al., 2015). Scientific articles are often restricted by the copyright of the publisher, and security risks, such as leakage, need to be prevented. As time passes, physical aging of the storage medium may also lead to data loss and difficulties in recovery. The scientific community's demand for research reproducibility creates additional challenges for data openness and sharing (Fecher et al., 2015). Finally, creating a clear and concise data file will also make storing bibliographic data more convenient.

Data sharing and data citations

Data sharing and data citations establish an interactive relationship between source and flow, where the source is the active distribution of data achievements to the scientific community by data creators, and the flow is the data citations. This relationship is the final stage of the scientometric life cycle and promotes the development of an open science culture. To this end, one of the main objectives of standardizing data storage is to support data sharing, while formal data citations honor data producers and, at the same time, advance data flows as creative elements of a scientific system. However, the significance of data sharing and citations has not been emphasized sufficiently. From the standpoint of responsible scientometric research, this section offers advice to the authors.

Share data via a variety of channels

Recently, many reputable journals have started requesting authors to disclose data, models, codes, and other pertinent attachment materials when publishing papers. To some extent, this ensures that research can be repeated and validated (Wilkinson et al., 2016). However, because most papers where the data are self-stored only provide access via an author's email address, the accessibility and usefulness of the data are limited. Consequently, we advise the authors to upload as much data as possible to a trustworthy platform. For instance, Mendeley Data, a platform for managing and sharing scientific research data created by Elsevier in 2015, offers a complete solution for data sharing, including data upload, data release, data storage, and data access (Bornmann & Haunschild, 2017); Zenodo, an open repository for all scholarships, enables researchers from all disciplines to share and preserve their research outputs, regardless of size or format. Free to upload and free to access, Zenodo makes scientific outputs of all kinds citable, shareable, and discoverable for the long term. Additionally, researchers can submit their data results for publication in peer-reviewed data journals, such as Scientific Data, which is part of the Nature Publishing Group. This is an important way to strike a balance between data sharing and performance acknowledgments. To encourage the exchange of high-quality data among peers, scientometric researchers should also try to share scientific research data more voluntarily, either through extensive collaboration or signing a data-sharing agreement. However, when someone shares the data, attention should be paid to the license under which data are shared (e.g., CC-BY or CC0), and they should not share data they are not allowed to share (e.g., raw data from Web of Science). To be effective in significantly increasing data availability, data-sharing policies should prescribe mechanisms for sharing that ensure reliable and long-term access to data (Federer et al., 2018), so including a data availability statement in research articles provides some potential solutions.

Formally cite open-access research data

Data citations generally refer to the practice of an author in identifying the source of the data used via a reference, footnote, or text note. The data cited include not only the sets of data that are crucial to the study but also the factual data that explain the background of the investigation and any additional data in the literature that attests

to the strength of the study's findings. Some studies report that the average number of data citations in scientific literature is low, but the proportion of informal citations is high (Park et al., 2018). Informal citations typically describe data sources in the context of tables, acknowledgments, or other parts of a paper. However, a standardized data index such as the Data Citation Index (DCI) cannot fully index informal citation records (Park & Wolfram, 2017, 2019). Consequently, we advise the authors to properly cite their data, following the format of a traditional literature citation. DataCite, which is working to standardize data citation practices and provide DOIs to datasets, is a guide that authors can use. According to the DataCite standard, a data citation must, at the very least, list the data's title, creators, publisher, DOIs, and year of publication (Robinson-García et al., 2016). The URL of the original source website is required if the data does not have a DOI. At the same time, we advise scientometric researchers to actively and consciously retrieve, cite, and analyze data records in DCI, Zenodo (Andrea et al., 2021), or another data resource platform to promote the development of such services.

Discussion and Conclusion

The scientometric framework proposed in this study systematically addresses key methodological challenges across the data lifecycle, from source selection to analytical interpretation. By deconstructing each operational phase, we revealed how technical decisions fundamentally shape research validity. While this work provides structured recommendations for standardization, it is critical to recognize that rigorous data practices alone cannot ensure responsible metrics in evaluation systems. The principles advocated in the San Francisco Declaration on Research Assessment (DORA) (DORA Group, 2013) and Leiden Manifesto (Hicks et al., 2015) remain essential complements, particularly regarding the contextualized use of metrics in research evaluation and their policy implications.

Moving forward, the scientometric field should treat methodological rigor and responsible metric use as interdependent requirements. Our phased framework provides scaffolding for reproducible research, but its implementation should actively engage with ongoing evaluation reform efforts, such as the Coalition for Advancing Research Assessment (CoARA). By explicitly addressing FRAME themes—such as enhancing transparency in evaluation processes, fostering a balanced use of bibliometric and qualitative insights, and promoting stakeholder engagement—this framework supports the development of more equitable and context-sensitive evaluation systems.

Additionally, the framework encourages the adoption of open science practices, such as sharing data and methodologies, to further enhance transparency and accountability. This aligns with the FRAME theme of promoting openness and trust in research evaluation. Only through this dual focus—optimizing technical processes while embedding ethical guidelines—can scientometrics fulfill its potential as a robust, socially accountable science of science. By aligning with initiatives like the Framework for Responsible Assessment Metrics in Research (FRAME), this study underscores the importance of transparency and accountability in shaping future evaluation systems. Such integration ensures that scientometric practices not only

advance methodological rigor but also contribute to a fairer and more sustainable research ecosystem.

Finally, the framework highlights the need for continuous dialogue between researchers, policymakers, and evaluators to ensure that bibliometric indicators are used in a way that aligns with societal values and research priorities. This participatory approach, rooted in the FRAME principle of inclusivity, further strengthens the framework's relevance to responsible research assessment.

Acknowledgments

We thank Alan L. Porter (Georgia Institute of Technology), Jiang Li (Nanjing University), Liming Liang (Henan Normal University), Lin Zhang (Wuhan University), Ludo Waltman (Leiden University), Ronald Rousseau (Antwerp University), Xianwen Wang (Dalian University of Technology), and Yishan Wu (Chinese Academy of Science and Technology for Development), for their valuable comments on this study.

References

Abdulhayoglu, M. A., & Thijs, B. (2017). Use of ResearchGate and Google CSE for author name disambiguation. *Scientometrics*, *111*(3), 1965-1985. https://doi.org/10.1007/s11192-017-2341-y

Andrea, S.-C., Nicolas, R.-G., van Thed, L., & Rodrigo, C. (2021). Exploring the relevance of ORCID as a source of study of data sharing activities at the individual-level: a methodological discussion. *Scientometrics*, *126*(8), 7149-7165. https://doi.org/10.1007/s11192-021-04043-5

Arora, S. K., Porter, A. L., Youtie, J., & Shapira, P. (2013). Capturing new developments in an emerging technology: an updated search strategy for identifying nanotechnology research outputs. *Scientometrics*, *95*(1), 351-370. https://doi.org/10.1007/s11192-012-0903-6

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79-80*, 3-15. https://doi.org/https://doi.org/10.1016/j.jpdc.2014.08.003

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century—A review. *Journal of Informetrics*, *2*(1), 1-52. https://doi.org/10.1016/j.joi.2007.11.001

Bornmann, L., Guns, R., Thelwall, M., & Wolfram, D. (2021). Which aspects of the Open Science agenda are most relevant to scientometric research and publishing? An opinion paper. *Quantitative Science Studies*, *2*(2), 438-453. https://doi.org/10.1162/qss_e_00121

Bornmann, L., & Haunschild, R. (2017). Measuring field-normalized impact of papers on specific societal groups: An altmetrics study based on Mendeley Data. *Research Evaluation*, *26*(3), 230-241. https://doi.org/10.1093/reseval/rvx005

Clarivate Analytics. (2020). *Web of Science Core Collection Help*. Retrieved 9 Oct 2022 from https://images.webofknowledge.com/images/help/WOS/hp_full_record.html

DORA Group. (2013). *The San Francisco Declaration on Research Assessment*. https://sfdora.org/

Egghe, L. (2015). Message from the retiring Editor-in-Chief. *Journal of Informetrics*, *9*(1), A1-A2. https://doi.org/10.1016/j.joi.2015.01.007

Fecher, B., Friesike, S., & Hebing, M. (2015). What Drives Academic Data Sharing? *Plos One*, *10*(2), e0118053. https://doi.org/10.1371/journal.pone.0118053

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *Plos One*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

Ferrara, A., & Salini, S. (2012). Ten challenges in modeling bibliographic data for bibliometric analysis. *Scientometrics*, *93*(3), 765-785. https://doi.org/10.1007/s11192-012-0810-x

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2016). The museum of errors/horrors in Scopus. *Journal of Informetrics*, *10*(1), 174-182. https://doi.org/10.1016/j.joi.2015.11.006

Fu, H.-Z., & Waltman, L. (2022). A large-scale bibliometric analysis of global climate change research between 2001 and 2018. *Climatic Change*, *170*(3), 36. https://doi.org/10.1007/s10584-022-03324-z

Gagolewski, M. (2011). Bibliometric impact assessment with R and the CITAN package. *Journal of Informetrics*, *5*(4), 678-692. https://doi.org/10.1016/j.joi.2011.06.006

Gusenbauer, M. (2022). Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. *Scientometrics*, *127*(5), 2683-2745. https://doi.org/10.1007/s11192-022-04289-7

Han, H., Yao, C., Fu, Y., Yu, Y., Zhang, Y., & Xu, S. (2017). Semantic fingerprints-based author name disambiguation in Chinese documents. *Scientometrics*, *111*(3), 1879-1896. https://doi.org/10.1007/s11192-017-2338-6

Harzing, A.-W. (2015). Health warning: might contain multiple personalities—the problem of homonyms in Thomson Reuters Essential Science Indicators. *Scientometrics*, *105*(3), 2259-2270. https://doi.org/10.1007/s11192-015-1699-y

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429-431.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569-16572. https://doi.org/10.1073/pnas.0507655102

Huang, Y., Zhu, D., Lv, Q., Porter, A. L., Robinson, D. K. R., & Wang, X. (2017). Early insights on the Emerging Sources Citation Index (ESCI): an overlay map-based bibliometric study. *Scientometrics*, *111*(3), 2041-2057. https://doi.org/10.1007/s11192-017-2349-3

Kim, J., & Diesner, J. (2015). The effect of data pre-processing on understanding the evolution of collaboration networks. *Journal of Informetrics*, *9*(1), 226-236. https://doi.org/10.1016/j.joi.2015.01.002

Kim, J., & Diesner, J. (2016). Distortive effects of initial-based name disambiguation on measurements of large-scale co-authorship networks. *Journal of the Association for Information Science and Technology*, *67*(6), 1446-1461. https://doi.org/10.1002/asi.23489

Lambert, N. (1991). Online searching of polymer patents: precision and recall. *Journal of Chemical Information and Computer Sciences*, *31*(4), 443-446. https://doi.org/10.1021/ci00004a002

Lin, C.-S., Huang, M.-H., & Chen, D.-Z. (2013). The influences of counting methods on university rankings based on paper count and citation count. *Journal of Informetrics*, *7*(3), 611-621. https://doi.org/10.1016/j.joi.2013.03.007

Liu, A.-Y., Fu, H.-Z., Li, S.-Y., & Guo, Y.-Q. (2014). Comments on "Global trends of solid waste research from 1997 to 2011 by using bibliometric analysis". *Scientometrics*, *98*(1), 767-774. https://doi.org/10.1007/s11192-013-1086-5

Liu, W. (2017). The changing role of non-English papers in scholarly communication: Evidence from Web of Science's three journal citation indexes. *Learned Publishing*, *30*(2), 115-123. https://doi.org/10.1002/leap.1089

Liu, W. (2019). The data source of this study is Web of Science Core Collection? Not enough. *Scientometrics*, *121*(3), 1815-1824. https://doi.org/10.1007/s11192-019-03238-1

Liu, W. (2021). Caveats for the use of Web of Science Core Collection in old literature retrieval and historical bibliometric analysis. *Technological Forecasting and Social Change*, *172*, 121023. https://doi.org/10.1016/j.techfore.2021.121023

Liu, W., Tang, L., & Hu, G. (2020). Funding information in Web of Science: an updated overview. *Scientometrics*, *122*(3), 1509-1524. https://doi.org/10.1007/s11192-020-03362-3

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871-906. https://doi.org/10.1007/s11192-020-03690-4

Newman, N. C., Porter, A. L., Newman, D., Trumbach, C. C., & Bolan, S. D. (2014). Comparing methods to extract technical content for technological intelligence. *Journal of Engineering and Technology Management*, *32*, 97-109. https://doi.org/10.1016/j.jengtecman.2013.09.001

Park, H., & Wolfram, D. (2017). An examination of research data sharing and re-use: implications for data citation practice. *Scientometrics*, *111*(1), 443-461. https://doi.org/10.1007/s11192-017-2240-2

Park, H., & Wolfram, D. (2019). Research software citation in the Data Citation Index: Current practices and implications for research software sharing and reuse. *Journal of Informetrics*, *13*(2), 574-582. https://doi.org/10.1016/j.joi.2019.03.005

Park, H., You, S., & Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, *69*(11), 1346-1354. https://doi.org/10.1002/asi.24049

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Pech, G., & Delgado, C. (2020). Assessing the publication impact using citation data from both Scopus and WoS databases: an approach validated in 15 research fields. *Scientometrics*, *125*(2), 909-924. https://doi.org/10.1007/s11192-020-03660-w

Porter, S. J. (2022). Measuring research information citizenship across ORCID practice. *Frontiers in Research Metrics and Analytics*, *7*, 779097. https://doi.org/10.3389/frma.2022.779097

Pranckutė, R. (2021). Web of Science (WoS) and Scopus: The Titans of Bibliographic Information in Today's Academic World. *Publications*, *9*(1), 12. https://www.mdpi.com/2304-6775/9/1/12

Robinson-García, N., Jiménez-Contreras, E., & Torres-Salinas, D. (2016). Analyzing data citation practices using the data citation index. *Journal of the Association for Information Science and Technology*, *67*(12), 2964-2975. https://doi.org/10.1002/asi.23529

Sivertsen, G., Rousseau, R., & Zhang, L. (2019). Measuring scientific contributions with modified fractional counting. *Journal of Informetrics*, *13*(2), 679-694. https://doi.org/10.1016/j.joi.2019.03.010

Strotmann, A., & Zhao, D. (2012). Author name disambiguation: What difference does it make in author-based citation analysis? *Journal of the American Society for Information Science and Technology*, *63*(9), 1820-1833. https://doi.org/10.1002/asi.22695

Velden, T., Hinze, S., Scharnhorst, A., Schneider, J. W., & Waltman, L. (2018). Exploration of reproducibility issues in scientometric research Part 2: Conceptual reproducibility. *arXiv preprint arXiv:1804.05026*.

Waltman, L., Hinze, S., Scharnhorst, A., Schneider, J. W., & Velden, T. (2018). Exploration of reproducibility issues in scientometric research Part 1: Direct reproducibility. *arXiv preprint arXiv:1804.05024*.

Waltman, L., Larivière, V., Milojević, S., & Sugimoto, C. R. (2020). Opening science: The rebirth of a scholarly journal. *Quantitative Science Studies*, *1*(1), 1-3. https://doi.org/10.1162/qss_e_00025

Waltman, L., & van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, *9*(4), 872-894. https://doi.org/10.1016/j.joi.2015.08.001

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship.
Scientific Data, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wilson, D. L., Tolson, J., Churchward, T. J., Melehan, K., O'Donoghue, F. J., & Ruehland, W. R. (2022). Exclusion of EEG-based arousals in wake epochs of polysomnography leads to underestimation of the arousal index. *Journal of Clinical Sleep Medicine*, *18*(5), 1385-1393. https://doi.org/10.5664/jcsm.9878

Youtie, J., Carley, S., Porter, A. L., & Shapira, P. (2017). Tracking researchers and their outputs: new insights from ORCIDs. *Scientometrics*, *113*(1), 437-453. https://doi.org/10.1007/s11192-017-2473-0

Zhang, J., Yu, Q., Zheng, F., Long, C., Lu, Z., & Duan, Z. (2016). Comparing keywords plus of WOS and author keywords: A case study of patient adherence research. *Journal of the Association for Information Science and Technology*, *67*(4), 967-972. https://doi.org/10.1002/asi.23437

Zhang, Y., Porter, A. L., Hu, Z., Guo, Y., & Newman, N. C. (2014). "Term clumping" for technical intelligence: A case study on dye-sensitized solar cells.
Technological Forecasting and Social Change, *85*,
26-39. https://doi.org/10.1016/j.techfore.2013.12.019

Zhu, J., & Liu, W. (2020). A tale of two databases: the use of Web of Science and Scopus in academic papers.
Scientometrics, *123*(1), 321-335. https://doi.org/10.1007/s11192-020-03387-8

https://doi.org/10.51408/issi2025_155

Responsible Metrics for The Assessment of Research Organizations

Gunnar Sivertsen[1], Lin Zhang[2], Alex Rushforth[3]

[1] *gunnar.sivertsen@nifu.no*
Nordic Institute for Studies in Innovation, Research and Education (NIFU), Oslo (Norway)

[2] *linzhang1117@whu.edu.cn*
School of Information Management, Wuhan University, Wuhan (China)
Department MSI, Centre for R&D Monitoring (ECOOM), KU Leuven, Leuven (Belgium)

[3] *a.d.rushforth@cwts.leidenuniv.nl*
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (Netherlands)

Research in progress: A contribution to the ISSI 2025 Special Track "A framework for the responsible use of bibliometrics in research evaluation" (FRAME).

Abstract

While the more recent initiatives for responsible research assessment, such as CoARA, primarily focus on two evaluation contexts, the assessment of researchers in their careers and of their proposals for project funding, *organizational-level assessment* is not as much discussed as it was ten years ago in relation to the Leiden Manifesto and the Metric Tide report. Organizational-level research assessment is traditionally where professional and science-based evaluative bibliometrics has contributed the most. This paper argues for a more systematic engagement with organizational research assessment within the ongoing reform of research assessment. Other factors are creating the opportunity to move in this direction: There is a general shift towards organizational evaluation in research assessment. There is also increasing availability of new diverse data on research activities produced by the global scientific publishing system. A third factor is an ongoing project organized the international RoRI institute in London which aims to create a global overview of national systems for assessment and funding of research organizations. Finally, we think it is time to appreciate that bibliometrics is not only involved in research assessment but also contributes to the science of science, thereby in turn laying the foundation for a proper understanding and assessment of institutionalized science.

Introduction

The Agreement on Reforming Research Assessment (CoARA, 2022) says that it addresses three contexts of evaluation, but it is focused mainly on the first two of them (Sivertsen & Rushforth, 2025):

1. Individual researchers as they apply for positions, promotions, or internal resources
2. Individual research proposals in applications for external funding
3. Research performing organisations and units

The engagement with primarily the first two contexts is understandable since the Agreement was developed in collaboration between the European University

Association and Science Europe with support from the European Commission. While the members of the EUA are directly responsible for assessments in the first context, members of Science Europe are directly responsible for assessments in the second context.

The third context of organizational assessment is addressed in the fourth of the Agreement's "Core commitments": "Avoid the use of rankings of research organisations in research assessment". The reference is to rankings provided by "external commercial companies" such as the *QS World University Rankings* and the *THE World University Rankings*. However, such rankings do not serve the purposes of the more widespread organizational assessments in research, which may be initiated by the research organizations themselves or by the funding governments as recurring national research assessment exercises. In our view, organizational research assessments across all countries also deserve attention from the perspective of responsible research assessment.

Bibliometric information at *aggregate levels* may be involved in organizational research assessment with motivations and indicators that differ from individual level assessments. Such use of bibliometrics did receive attention in earlier phases of the movement towards improved practices in research assessment, e.g., in the *Leiden Manifesto* (Hicks et al., 2015) and the *Metric Tide* report (Wilsdon et al., 2016), but it is so far overlooked by CoARA.

The third "Core commitment" of the Agreement does not provide much help in contexts of organizational research assessment: "Abandon inappropriate uses in research assessment of journal- and publication-based metrics, in particular inappropriate uses of Journal Impact Factor (JIF) and h-index". Both professional scientometricians and research assessment reform initiatives like CoARA agree the two named indicators are inappropriate for assessment researchers' CVs and proposals of their colleagues, or at an organizational level – and both share a concern about their widespread application in these very contexts. But beyond H-index, JIF, and commercial university rankings, the role of advanced bibliometrics in organizational assessments is ambiguous in ARRA. ARRA writes critically of "journal- and publication-based metrics", but seems to be unaware that this term comes close to a definition of *bibliometrics*[1], a term never used in the document. A blog more recently published by the CoARA steering board appears to acknowledge that bibliometrics can play a role at the organizational level, but currently, this issue remains an underdeveloped part of the CoARA project.

In our view, a set of appropriate (for the type and profile of the institution and the purpose of the evaluation) advanced science-based bibliometric indicators can be helpful, implying that "publication-based metrics" (=bibliometrics) should not be abandoned. There is a long tradition in science-based bibliometrics for serving organizational level assessments and statistics with advanced indicators while at the same time discussing their limitations, particularly at the individual level. Newcomers to the field are trained to understand the importance of the level of

[1] "Bibliometrics denotes the quantitative study of publications, citations, and related surrogate measures in scholarly communication." (Broadus, 1987)

analysis, as shown in the illustration[2] which was developed by Wolfgang Glänzel for use at the *European Summer School for Scientometrics (ESSS)*.
In our view, there is a need to develop a shared understanding with CoARA of the strengths and limitations of bibliometrics in research assessment. We suggest that one way forward could be to focus more on systems, practices, and indicators for organizational level research assessment. Our contribution in this paper is to discuss *four reasons* for giving organizational research assessment more attention in dialogue with CoARA.

1) A trend towards organizational assessment

There is an increasing interest worldwide in research performance, not just as a sum of individual achievements, but as an *institutional responsibility*. CoARA is an expression of – and has a role in – a historical shift of evaluation paradigms during three decades from 1) research assessment among experts within disciplines via 2) excellence-orientation (to favour the best in competition across disciplines) towards responsible research assessment, which is more focused on assessing the conditions for performing good research by broadening the empirical basis for the assessment and including societal relevance and challenges. The historical shifts are observable in the Research Excellence Framework (REF) in the UK, particularly in the revised purposes and requirements of the next REF 2029 (https://2029.ref.ac.uk/). Compared to the two recent rounds in 2014 and 2021 and the earlier rounds of the preceding RAE since 1986, one main trait is evident: The research assessment in the UK is downscaling the role of 1) quality assessment of individual outputs of research, and, on the other hand, extending the role of 2) the assessment of the research performing organization as such, which will now be named People, Culture and Environment (PCE). In addition, 3) the societal impact of research has been included in the UK since 2014.

2) A new global overview of national research assessment and funding systems

The trend described above has become observable within the framework of the AGORRA project (A Global Observatory of Responsible Research Assessment), from which the international Research on Research Institute (RoRI) in London is currently publishing a report titled *A New Typology of National Research Assessment and Funding Systems: Continuity, Change, and Contestation Across Thirteen Countries* (RoRI Working Paper No. 15). We are involved in this project. The report aims to establish an online global monitor of national research assessment and funding systems, and to expand the coverage to more countries. Currently, the study includes expert contributions from thirteen countries: Argentina, Australia, Brazil, Chile, China, Colombia, India, Italy, Mexico, the Netherlands, Norway, Poland, and the United Kingdom. One of the key aspects examined is the use of bibliometric

[2] Illustration *The weight of qualitative (peer evaluation) and quantitative (bibliometrics) methods as function of the aggregation level* by Wolfgang Glänzel, from the presentation *Thoughts and Facts on Bibliometric Indicators in the Light of New Challenges in Their Applications* at the European Summer School for Scientometrics (ESSS).

indicators within each system, as some countries employ multiple frameworks. The project also highlights that organizational-level research assessment is widely practiced across all countries and provides a typology of the systems the provides for cross-national comparisons. Despite the significant variability in research assessment and funding systems, shaped by unique historical, cultural, and policy contexts, few countries rely solely on indicators for organizational-level assessment. A comprehensive understanding of these diverse frameworks is crucial for developing responsible and effective evaluation practices. Observing the national differences underscore the need to contextualize research assessment within national priorities and institutional missions while at the same time fostering a global dialogue on responsible research assessment and opening up for mutual learning. Engaging with such projects can enhance the development of fair and effective evaluation frameworks that respect the distinctive characteristics of different research environments. Furthermore, collaboration with this type of projects could strengthen the dialogue between CoARA and the international bibliometrics community, promoting a more inclusive and informed approach to research evaluation.

3) New sources of data are emerging in the scientific publishing system

Parallel to the trend and the new opportunity described above, the digital universe of scientific publishing has developed quickly towards creating new types of data that may extend the range of indicators for organizational assessment. The CoARA Agreement lists several items that should be assessed in addition to publications: data, software, models, methods, theories, algorithms, protocols, and exhibitions. All of them are now publishable within a publication, in an appendix, or in linked documents. In fact, all indicators of responsible research practices published with the *Hong Kong Principles for assessing researchers* (Moher et al. 2020) may now be represented in a scientific publication or by indicators derived from it. The Agreement also says: "Value a range of other contributions to responsible research and scholarly activity, such as peer review for grants and publications, mentoring, outreach, and knowledge exchange". Again, data sources and indicators for such activities are being developed within the scientific publishing system. Examples are those mentioned in the Annex of ARRA: Open science badges; Publons, ORCID, open peer review; CRediT; reporting guidelines (e.g. EQUATOR Network); and altmetrics (Altmetrics, PlumX). Most of these data sources have already been introduced in studies published by the main international journals for bibliometrics. The scientific attention thereby given to the new types of data may in turn be helpful for the reform of research assessment.

4) The science of science is needed

Finally, to mention not only recent trends, but an important long tradition as well: Bibliometrics is not only there to serve research assessment, and it is not a recent invention by commercial suppliers. Bibliometrics represents hundred years of contributions to the science of science. These contributions provide insights that are necessary to understand science as an organized activity in society and thereby a

foundation for responsible assessments of research that are appropriate for the different profiles and purposes of its organizations.

Conclusions

The responsible use of bibliometric indicators in research assessment should not be limited to individual researchers and project-level evaluations. The increasing institutional responsibility for research performance, the growing availability of diverse research output data, and global trends in assessment frameworks underscore the need for bibliometric approaches at the organizational level. Dialogues with CoARA and similar initiatives can help bridge the gap between bibliometric expertise and policy discussions, ensuring that assessments align with principles of responsible research evaluation while maintaining methodological rigor. Future efforts should focus on refining indicators, improving transparency, and fostering collaboration between the bibliometric community and research policy stakeholders.

References

Broadus, R. N. (1987). Toward a definition of "bibliometrics". *Scientometrics*, 12, 373-379. https://doi.org/10.1007/BF02016680.

CoARA. (2022). Agreement on Reform of Research Assessment. https://coara.eu/agreement/the-agreement-full-text/

Hicks, D., Wouters, P., Waltman, L. et al. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520, 429–431.

Moher, D., et al. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. *PLoS Biology*, 18(7): e3000737.

Sivertsen, G., & Rushforth, A. (2025). The Ongoing Reform of Research Assessment. In: Sivertsen, G., Langfeldt, L. (eds) *Challenges in Research Policy*. SpringerBriefs in Political Science. Springer, Cham. https://doi.org/10.1007/978-3-031-69580-3_7.

Wilsdon, J. et al. (2015) *The Metric Tide: The Independent Review of the Role of Metrics in Research Assessment and Management*. https://doi.org/10.13140/RG.2.1.4929.1363.

https://doi.org/10.51408/issi2025_156

Responsible Uses of Large Language Models for Research Evaluation

Mike Thelwall

m.a.thelwall@sheffield.ac.uk
Information School, University of Sheffield (UK)

Abstract

Although research evaluators and scientometricians have promoted the message of responsible bibliometrics through initiatives like the Leiden Manifesto, these do not mention Large Language Models (LLMs). LLMs can now make useful quality predictions for journal articles, giving values that correlate more strongly with expert judgements than do citation-based indicators in most fields. This has created the possibility that they could supplement or even replace citation-based indicators for some applications. As tested so far, LLMs predict the quality rating that a human expert would give a paper. They do this by reading the quality level descriptions and then processing the article title and abstract. This raises multiple new issues in comparison to the Leiden Manifesto. First, authors might try to trick LLMs into giving high scores by crafting LLM-friendly abstracts. Second, LLM models incorporate billions of parameters, so their scores are opaque. Third, it is not clear how LLMs work in terms of the main influences on their scores, so their biases are unknown. Fourth, whilst citations reflect tangible and permanent contributions to the scientific record, albeit of variable value, LLM-based predictions do not clearly link to progress. Fifth, LLM scores are ephemeral in the sense that newer LLMs may give substantially different scores and rankings.

Introduction

Research evaluation is often used to support decision making. For example, job applicants may be judged on the quality of their work, departmental funding might be dependent on positive research quality or volume evaluations, and national policy may be informed by estimates of the areas in which the country appears strong relative to its competitors. In these cases, there are winners and losers, assuming that there are finite resources to allocate. Thus, it is not only important to ensure that the research evaluations are as accurate as possible, but also that they are not biased in a way that would undermine the system. These two considerations do not always fully align: if the research evaluation method that is the most accurate overall also has a substantial bias against a particular group (e.g., women, ethnic minorities, applied researchers), then it might not be acceptable for reasons of social equity or national policy.

A research evaluation approach might also be ruled irresponsible if it generates perverse incentives. Whenever people are evaluated and know the evaluation method, it is natural for some to target the method rather than the underlying goal, potentially generating unwanted outcomes. For example, if academics are evaluated on the number of articles they produce then they might divert some of their effort into publishing smaller and possibly weaker articles at the expense of books, chapters, conference papers, and long articles (Aagaard, 2015).

A third issue is transparency: the ability of those evaluated, or affected by an evaluation, to see the details of the mechanism used to evaluate them. This may give confidence in the evaluation system and may improve it if errors can be identified and corrected. In practice, transparency is always partial. The most transparent system might be citation-based indicators from OpenAlex since it publishes the source code of all the algorithms it uses (Priem et al., 2022). This is still not full transparency because its citation counts are based on citations made by millions of individual scientists behind closed doors. A deliberate lack of transparency is also common in research: authors are rarely told the identities of the reviewers rejecting their paper or giving a low score to their grant (i.e., single/double blind peer review), and some decisions are made without any explanation or rationale.

In the light of these considerations, it seems reasonable to suggest that research evaluations ought to be as responsible as possible, in the sense of minimising the above risks as far as is practical in the context of the goals and resources of the evaluation. It also seems like good practice to be honest about the extent to which any problems remain. The rest of this paper briefly summarises some responsible research evaluation initiatives for bibliometrics and then focuses on the considerations that are relevant to the use of large language models to support research evaluation.

Responsible bibliometrics

Perhaps the most well-known responsible bibliometrics initiative is the Leiden Manifesto (Hicks et al., 2025). Its ten principles are:

1. Quantitative evaluation should support qualitative, expert assessment.
2. Measure performance against the research missions of the institution, group, or researcher.
3. Protect excellence in locally relevant research.
4. Keep data collection and analytical processes open, transparent, and simple.
5. Allow those evaluated to verify data and analysis.
6. Account for variation by field in publication and citation practices.
7. Base assessment of individual researchers on a qualitative judgement of their portfolio.
8. Avoid misplaced concreteness and false precision.
9. Recognize the systemic effects of assessment and indicators.
10. Scrutinize indicators regularly and update them (Hicks et al., 2025).

Overall, the Leiden Manifesto goal is to reduce the chance that bibliometrics are used unwisely for research evaluation. The Metric Tide (Wilsdon et al., 2013) is similar but more UK focused.

There are other prominent initiatives against inappropriate uses of specific types of indicators as part of a wider movement for assessment reform (Rushforth, 2025; Rushforth & Hammarfelt, 2023). The San Francisco Declaration on Research Assessment (DORA; sfdora.org) campaigns against overuse of journal-based

indicators in the belief that research evaluation should focus on articles rather than publication venues and that focusing too much on journals creates a perverse incentive that is unhealthy to the diversity of scientific publishing. This follows many years of criticisms of article-level citation-based indicators and journal impact factors (e.g., MacRoberts & MacRoberts, 2018; Rushforth & de Rijcke, 2015; Seglen, 1998).

In parallel, More Than Our Rank (inorms.net/more-than-our-rank) campaigns against reliance on league tables of universities. Focusing on league tables can cause perverse incentives, such as hiring academics for their citation rates or prizes rather than their ability to support the university goals (if different). These league tables usually either rely on citation rates or have them as an important component but the other methods used are also flawed. For example, reputational surveys favour older and larger institutions because more academics will know them, giving them a larger potential voter base (Gadd, 2020; Vernon et al., 2018).

As these examples show, specific problems with bibliometrics and related research evaluation methods have given rise to initiatives to combat them. With the rise of Artificial Intelligence (AI) support for research evaluation, potential new problems must also be considered.

Research evaluation applications of LLMs

There have been many attempts to introduce AI in the form of traditional machine learning into research evaluation, such as to predict long term citation rates for recently published articles (Ma et al., 2021), but they do not seem to have led to any practical applications. The situation seems set to change with the rise of Large Language Models (LLMs), which have some capability to follow human instructions for text processing tasks (Ouyang et al., 2022) and perform well in many cases (Kocoń et al., 2023). In this context, early evidence suggests that they have a technical capability to challenge bibliometrics as the most accurate scientific research quality indicator. Specifically, small-scale studies have shown that research quality judgments by ChatGPT for submitted or published articles correlate positively with private human judgements or scores (Saad et al., 2024; Thelwall, 2024) and in some cases for public scores (Zhou et al., 2024; Thelwall & Yaghi, 2025). In addition, a large-scale study has suggested that ChatGPT quality predictions may correlate more strongly than citation-based indicators with research quality scores for most academic fields (Thelwall & Yaghi, 2024; Thelwall et al., 2025; Thelwall, 2025). Since this accuracy creates the possibility that LLMs may complement or replace citation-based indicators in the future, it is important to consider how this might impact on considerations for responsible indicators.

Possible Applications of LLM research quality scores

In theory, LLMs could be used for most evaluation roles that citation-based indicators currently fill. The main current exception is that some citation indicators are network-based, such as evidence of the countries in which a nation's or journal's citations mainly originate (Schubert & Glänzel, 2006; Zhang et al., 2009). This section discusses some likely research evaluation applications of LLMs.

Support for article-level expert quality ratings

Individual articles sometimes need to be assessed or scored for quality for job-related reasons (appointments, tenure, promotion), impacting academic careers. Currently these evaluations might be formal (e.g., asking experts to read and score articles) or informal (e.g., forming a quick impression of a candidate's research strengths by browsing their CV). Heuristics seem likely to be used for quick informal evaluations and those made by people that are not experts on the candidate's topics. These might typically include journal reputation, journal citation rates, and article citation counts. Overinterpreting the results is a common cause of concern (Rushforth & De Rijcke, 2024). LLMs could be used in a similar way, in theory. In practice, it seems unlikely that LLMs would often be used well in this role since they need some knowledge to set up and their scores need to be rescaled from multiple submissions to be meaningful (Thelwall, 2024). Thus, LLMs are currently more of a threat than a benefit in this role, until a system exists that would generate meaningful score predictions (e.g., scaled to align with human judgement).

LLMs might currently be most useful for large scale formal evaluations like the UK Research Excellence Framework (REF), which individually scores over 100,000 journal articles and uses citation-based indicators in a minor role for some health and physical sciences fields and economics. The citation-based indicators are carefully selected and curated, and the same could be achieved for LLMs scores. They might also be useful for a wider range of fields than bibliometrics, including some where they had a stronger correlation with expert judgements than do citation rates (Yaghi & Thelwall, 2024).

Departmental-level evaluations

In some situations, departments are evaluated as a whole, either individually by benchmarking them against other similar departments or as part of a national evaluation of all departments of a given type. Here, it seems plausible that average LLM scores could be calculated as an additional indicator to citation-based indicators. It would be interesting to see if this helped any department type. Again, LLMs might be used across a wider range of fields than citations currently are.

National and international comparisons

In theory, citation-based bibliometric analyses of national strengths and weaknesses, as included in periodic reports by or for governments (e.g., Science, Research and Innovation Performance of the EU) could be supplemented by a section on LLM scores, potentially expanding indicator coverage beyond fields for which citation-base indicators have the most value.

JIFs

Average LLM scores for articles published by a journal can be calculated as an alternative to the average citation rates of the Journal Impact Factor (JIF) and similar formulae. The results from the two approaches correlate positively and moderately or strongly, depending on the field. Moreover, the LLM version may be fairer to

journals that attract relatively few citations because the citing journals are not included in a citation index (Thelwall & Kousha, 2025).
An advantage of LLM-based journal quality indicators is that they could be based on the most recent year of published articles, rather than older articles, as currently used for all well-known citation-based journal impact indicators. This would make the results more current. A potential disadvantage is that if LLM-based journal ranking becomes common then publishers and editors may attempt to at least partly target the journal's formatting requirements or style guidelines towards LLM-friendly elements. It is not clear what this would entail.

Threats to responsible uses of LLMs

This section discusses three of the Leiden Manifesto's most relevant aspects for LLMs.

Perverse incentives

Since perverse incentives occur by people targeting indicators rather than the underlying goal, the logical LLM perverse incentive is for authors to craft articles for high LLM scores rather than for communicating their research accurately and clearly for a human audience. This could be achieved by entering an article into an LLM and asking it to make suggestions to make it more likely to achieve a high score. This might involve exaggerating the importance of the findings to make the research more like a press release.
Whilst it seems likely that authors would attempt to do this, wasting their time on an unproductive activity, it is not clear that it would work to any great extent. Articles go through peer review and this guards against unsupported claims, so authors enlisting LLM help might find their work more likely to be rejected. Moreover, there are many LLMs, they have different strengths, and they evolve over time so it is not clear that crafting an LLM-friendly article would work even if it passed peer review. In addition, if the practice was suspected then evaluators might try to detect and penalise LLM-supported articles.
Thus, overall, it seems reasonably likely that the main perverse incentive is for authors to waste time on creating LLM-friendly work rather than that these attempts would succeed and lower the accuracy of LLM-based evaluations.

Transparency: Opaque LLM scores

LLMs have arguably the same transparency issues as peer review. In the same way that we can't see the processes going on inside a reviewer's brain when they cogitate over what they have read and experienced, turning their knowledge into a score/judgement and report, we also can't follow the numerous weights (typically above 7 billion even for the smallest model) within an LLM leading to its score and justification. In theory, an LLM could have more transparent inputs than a human reviewer in the sense that it could be trained on a known corpus of work (e.g., everything in OpenAlex), and LLM algorithms are certainly more understandable than human brains, at least in their overall architecture. These seem to be minor differences, however, given the overall complexity of even the smallest LLM. In

contrast, bibliometrics seem to be more transparent. For example, citation-based indicators from OpenAlex are relatively transparent, as argued above. Here the main opaqueness, the citing author decisions, is perhaps less important because each decision is relatively minor if many citations are counted for an evaluation.

Biases: Unknown LLM score influences and biases

Since LLM evaluations are relatively new, little is known about their biases. In contrast, some bibliometrics have been shown to have gender biases (e.g., career citations) and most have international biases, and there may also be institutional, reputational and interdisciplinary disparities (e.g., Paris et al., 1998; Schisterman et al., 2017). For ChatGPT-based evaluations it is known that some fields get substantially higher average scores than others (Thelwall & Yaghi, 2024), but little else is known about any other biases.

Research into AI biases in other contexts has shown that apparently objective mathematical algorithms can be biased if they are fed with unbalanced data or misleading assumptions by their engineers. They can also generate new biases as an unintended side effect of their data and algorithm (Akter et al., 2022; Baeza-Yates, 2016). Thus, it is reasonable to expect that LLMs will have learned biases from their inputs and may also have generated new ones. The extent and nature of these is not known, however. It is therefore an important due diligence step for researchers to test LLM scores for the most likely and worrying types of disparity.

New LLMs Irresponsibility Dimensions?

As shown above, responsible uses of LLMs should consider the same issues as for bibliometrics. There are additional considerations that do not apply to citation-based indicators, and some are discussed here.

Ephemerality and variation between LLMs

An important difference between citations and LLMs currently seems to be that citations are tangible and verifiable, whereas LLM judgements are not. In particular, an author judged to have 20 excellent papers by one LLM might next year be judged to have only five by a different LLM or an improved version of the same one. Whilst this perhaps mirrors the peer opinion situation in the sense that a person's work might go out of fashion, it must be demoralising to know that research achievements can disappear suddenly due to an algorithm change. This might reduce confidence in the research evaluation system, if used for individual academics.

There are at least three ways to address this issue. First, research into the stability of LLM scores might give reassurance that wholescale score shifts, as hypothesised above, are unlikely. Second, only aggregate scores across multiple articles might be used for evaluations, reducing the impact of changes for individual articles. Third, long term evaluation processes might build in stability, such as by fixing a score at a given point in time or altering scores primarily by adding new evaluations rather than replacing old evaluations.

Alignment of prompts with evaluation goals

Unlike citations, LLM prompts might be tailored to the goals of a research evaluation. For example, if the goal is the generic one of assessing research quality, then the prompt might ask the LLM to assess an article for the three core dimensions of rigour, significance and originality (Langfeldt et al., 2020), perhaps tailoring the definitions of these to a local context or with local examples. Responsible evaluators would have the option to tailor their prompts to more specific goals, however, such as value to the national economy or support for United Nations Development Goals. Of course, tailoring the prompts to a particular goal does not mean that the LLM will be capable of responding appropriately.

Lack of connection to research progress

Another dimension of uncertainty for LLM scores is that they do not have a direct theorised connection to research progress. For citations, Merton's (1973) theory posits that citations are scholarly acknowledgements of prior work that has aided the creation of new research. This is an oversimplification since the selection of work to cite is subjective with influential prior work often remaining uncited (e.g., obliteration by incorporation: McCain, 2011) and work without influence being cited (e.g., for background context). Nevertheless, it is still possible to claim that in many fields some citations reflect influence, and the rest are noise, with the latter tending to disappear at a sufficiently high level of aggregation (van Raan, 1998). There does not seem to be a way to mitigate the lack of a tangible connection to research progress for LLM evaluations, although it is the same as for expert opinions.

Cost

The relative costs of LLMs and bibliometric indicators are not yet clear. If wide uptake is to be achieved, LLM scores might need to be offered by citation index providers. These would be able to share the costs of the LLM queries or processing across all users. Citation-based indicators currently (March 2025) have two advantages: there are no providers of LLM-based academic scores and OpenAlex is a free source of citation-based indicators. Of course, the cost of LLM scores includes the personnel costs associated with the skills needed to generate the scores as well as the computing costs.

Summary

As argued above, issues relevant to the responsible use of LLM-based quality scores are partly the same as for bibliometrics and partly different, with some new considerations. Returning to the Leiden Manifesto (Hicks et al., 2025), the LLM adjustments can be summarised as follows.

1. *Quantitative evaluation should support qualitative, expert assessment*. This is the same for LLMs, even though they mimic human peer review more than do citations.
2. *Measure performance against the research missions of the institution, group, or researcher*. The same for LLMs.

3. *Protect excellence in locally relevant research*. The same for LLMs. They may have more capacity to do this than citation-based indicators since LLM prompts could be explicitly tailored to local goals, needs and concepts of research quality.
4. *Keep data collection and analytical processes open, transparent, and simple*. Current major LLMs fail this, as discussed above, with the partial exception that their algorithm architectures are known, a minor advantage over human brains.
5. *Allow those evaluated to verify data and analysis*. Current LLMs fail this because they do not publish their data sources, except that those analysed could replicate the actions of those obtaining the scores, if they publish their prompts and the identity of the LLM used. Because of the random parameters used in LLMs, they will not get the same results and might get substantially different results occasionally. This issue could be addressed by evaluators only using offline LLMs and publishing the random seed values, but this seems like a minor point.
6. *Account for variation by field in publication and citation practices*. Users of LLMs should consider variations between fields in the average LLM scores.
7. *Base assessment of individual researchers on a qualitative judgement of their portfolio*. This is the same for LLMs.
8. *Avoid misplaced concreteness and false precision*. This is the same for LLMs.
9. *Recognize the systemic effects of assessment and indicators*. This is also important for LLMs although the issues are different, as discussed above.
10. *Scrutinize indicators regularly and update them*. This seems likely to happen naturally for LLMs as new versions appear and existing ones evolve. New prompts should also be tested especially to align to local needs.

To these ten points, four additional suggestions could be incorporated for LLM-based scores.

11. *Design prompts to align with the research evaluation goals*.
12. *Consider the ephemerality of scores and differences between LLMs*.
13. *Consider the costs of generating LLM scores relative to bibliometric alternatives.*
14. *Accept that LLM scores are not direct evidence of contributions to science*.

References

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. *Science and Public Policy*, *42*(5), 725-737.

Akter, S., Dwivedi, Y. K., Sajib, S., Biswas, K., Bandara, R. J., & Michael, K. (2022). Algorithmic bias in machine learning-based marketing models. Journal of Business Research, 144, 201-216.

Baeza-Yates, R. (2016). Data and algorithmic bias in the web. In Proceedings of the 8th ACM Conference on Web Science. https://doi.org/10.1145/2908131.2908135

Gadd, E. (2020). University rankings need a rethink. Nature, 587(7835), 523-524.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. Nature, 520(7548), 429-431.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. Information Fusion, 99, 101861.

Langfeldt, L., Nedeva, M., Sörlin, S., & Thomas, D. A. (2020). Co-existing notions of research quality: A framework to study context-specific understandings of good research. Minerva, 58(1), 115-137.

Ma, A., Liu, Y., Xu, X., & Dong, T. (2021). A deep-learning based citation count prediction model with paper metadata semantic features. Scientometrics, 126(8), 6803-6823.

MacRoberts, M. H., & MacRoberts, B. R. (2018). The mismeasure of science: Citation analysis. Journal of the Association for Information Science and Technology, 69(3), 474-482.

McCain, K. W. (2011). Eponymy and obliteration by incorporation: The case of the "Nash Equilibrium". Journal of the American Society for Information Science and Technology, 62(7), 1412-1424.

Merton, R. K. (1973). The sociology of science: Theoretical and empirical investigations. University of Chicago Press.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

Paris, G., De Leo, G., Menozzi, P., & Gatto, M. (1998). Region-based citation bias in science. Nature, 396(6708), 210-210.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833.

Rushforth, A. (2025). Research Assessment Reform as Collective Action Problem: Contested Framings of Research System Transformation. Minerva, 1-21.

Rushforth, A., & de Rijcke, S. (2015). Accounting for impact? The journal impact factor and the making of biomedical research in the Netherlands. Minerva, 53(2), 117-139.

Rushforth, A., & De Rijcke, S. (2024). Practicing responsible research assessment: Qualitative study of faculty hiring, promotion, and tenure assessments in the United States. Research Evaluation, 33, rvae007.

Rushforth, A., & Hammarfelt, B. (2023). The rise of responsible metrics as a professional reform movement: A collective action frames account. Quantitative Science Studies, 4(4), 879-897.

Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., Vaishya, R., & Botchu, R. (2024). Exploring the potential of ChatGPT in the peer review process: An observational study. Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 18(2), 102946. https://doi.org/10.1016/j.dsx.2024.102946

Schisterman, E. F., Swanson, C. W., Lu, Y. L., & Mumford, S. L. (2017). The changing face of epidemiology: gender disparities in citations? Epidemiology, 28(2), 159-168.

Schubert, A., & Glänzel, W. (2006). Cross-national preference in co-authorship, references and citations. Scientometrics, 69, 409-428.

Seglen, P. O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. Acta Orthopaedica Scandinavica, 69(3), 224-229.

Thelwall, M., & Yaghi, A. (2024). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. arXiv preprint arXiv:2409.16695.

Thelwall, M. & Yaghi, A. (2025). Evaluating the predictive capacity of ChatGPT for academic peer review outcomes across multiple platforms. Scientometrics, to appear.

Thelwall, M., Jiang, X., & Bath, P. (2025). Estimating the quality of published medical research with ChatGPT. Information Processing & Management, 62(4), 104123. https://doi.org/10.1016/j.ipm.2025.104123

Thelwall, M., & Kousha, K. (2025). Journal Quality Factors from ChatGPT: More meaningful than Impact Factors? Journal of Data and Information Science. https://doi.org/10.2478/jdis-2025-0016

Thelwall, M. (2024). Can ChatGPT evaluate research quality? Journal of Data and Information Science, 9(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Thelwall, M. (2025). In which fields do ChatGPT 4o scores align better than citations with research quality? *arXiv preprint arXiv:2504.04464.*

van Raan, A. F. (1998). In matters of quantitative studies of science, the fault of theorists is offering too little and asking too much. Scientometrics, 43, 129-139.

Vernon, M. M., Balas, E. A., & Momani, S. (2018). Are university rankings useful to improve research? A systematic review. PloS One, 13(3), e0193762.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., & Johnson, B. (2015). The metric tide. Report of the independent review of the role of metrics in research assessment and management. https://www.ukri.org/wp-content/uploads/2021/12/RE-151221-TheMetricTideFullReportLitReview.pdf

Zhang, L., Glänzel, W., & Liang, L. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. Scientometrics, 81(3), 821-838.

Zhou, R., Chen, L., & Yu, K. (2024). Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 9340-9351).

https://doi.org/10.51408/issi2025_157

Mapping National Research That Targets Sustainable Development Goals: The Responsible Visualization of Openalex Data for Societal Impact Measurements of Research

Robin Haunschild[1], Lutz Bornmann[2]

[1]*R.Haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Information Service,
Heisenbergstr. 1, 70569 Stuttgart (Germany)

[2]*L.Bornmann@fkf.mpg.de, bornmann@gv.mpg.de*
Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society,
Hofgartenstr. 8, 80539 Munich (Germany)

Abstract

As one approach to a framework for the responsible use of bibliometrics in research evaluation, we proposed to use global maps using OpenAlex to highlight concepts (i.e., fields of research) where countries are particularly active to achieve United Nations sustainable development goals (UN SDGs). As first examples in this research-in-progress paper, we used the USA and Japan. As to be expected, we found that the USA is very active in many concepts to achieve the SDGs (since the USA is very research active in general). We revealed for Japan that the country has two increased areas of activity to achieve the SDGs: One area is in Medicine and the other in Chemistry and Material Sciences. Our SDG mapping approach combines multiple aspects of the responsible use of bibliometrics in research evaluation: (1) By focusing on SDG relevant research, we provide an innovative approach for measuring target-oriented the societal impact of research. (2) Our approach goes beyond using simple counting of publications or citations by using maps to display complex results. (3) The usage of OpenAlex and free statistics software makes our procedure transparent and reproducible.

Introduction

In recent years, some initiatives have been started with the goal of reforming the way research is assessed (Rushforth & Hammarfelt, 2023). The initiatives include the Leiden Manifesto (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015), the Metric Tide (Wilsdon et al., 2015), the Declaration on Research Assessment (DORA, https://sfdora.org/), and the Agreement on Reforming Research Assessment (CoARA, https://coara.eu). Whereas DORA focuses on reducing the use of journal-based citation impact indicators in research assessments, CoARA emphasizes the need for a more holistic approach to research evaluation (Thelwall, 2024). In this research-in-progress paper, we took up this call for a more holistic approach by introducing science maps visualizing national research that targets United Nations sustainable development goals (UN SDGs, https://sdgs.un.org/goals). The maps are intended to highlight the areas in which national research targets (worldwide) societal challenges. Most of the previous maps have focused on the visualization of traditional metrics, i.e., citation impact of publications.

In 2000, the UN established six Millennium Development Goals and in 2015, adopted the 2030 Agenda, which includes 17 interconnected SDGs. The Agenda

outlines an action plan for people, planet, and prosperity. At the Stockholm conference in 2022 (https://www.stockholm50.global), proposals were made to accelerate the achievement of the 2030 Agenda, focusing on SDGs for a healthy planet, social and economic progress, well-being, and resilience (Hernandez, Suazo López, & Domínguez Pacheco, 2022). It is one important goal of the science system to encompass societal products (outputs), societal use (societal references), and societal benefits (changes in society). It has been argued that society only reaps benefits from successful scientific studies when their results are converted into products (e.g., medications, diagnostic tools, machines, or devices) or services (e.g., government advising) (Bornmann, 2012, 2013). In recent years, some studies were published that have investigated whether scientific studies not only have societal impact but also specifically address SDGs (Ciarli, 2022; Purnell, 2022).

Using data from the OpenAlex database, we propose in this study overlay maps that visualize the national research that is especially active in worldwide SDG-relevant research. These overlay maps are visual tools used to represent the relationships and positions of national data within the worldwide scientific landscape. The maps overlay national data onto a base map that represents the entire science system. This helps to visualize how the national data fit into the larger context of scientific research. To demonstrate the overlay maps in this study, we present the maps for the USA and Japan.

Contribution of global overlay maps using OpenAlex to responsible bibliometric practices

In the development of the global overlay maps technique presented in this paper, we tried to follow the various guidelines for the responsible use of bibliometrics. The Leiden Manifesto (Hicks, et al., 2015) presents ten principles to guide research evaluation. The fourth principle suggests to use open data to foster transparency in research evaluation. CoARA also calls for the use of open datasets for and transparency in research evaluation. Since OpenAlex is openly available, we decided to use OpenAlex in order to follow both guidelines. The fourth principle of the Leiden Manifesto suggests that evaluation methodologies should be transparent. By laying our methodology out in this contribution, we also follow this principle. Adams, McVeigh, Pendlebury, and Szomszor (2019) argue for using profiles rather than metrics in research evaluation. Our proposal of using overlay maps to visualize contributions to reaching SDGs is one step in that direction. An earlier step into that direction was the introduction of beam plots for raw citations (Haunschild, Bornmann, & Adams, 2019). DORA suggests to consider a broad range of impact measures in research evaluation. CoARA also calls for "consideration of contributions to the research ecosystem, knowledge generation and scientific, technological, economic, cultural and societal impact" (CoARA, 2022). By providing a transparent methodology for the analysis of publications that targets SDGs, we extend the range of impact measures for research evaluation with the use of our maps.

Methods and data

Assignments of papers to SDGs

Assignments of papers to SDGs is made in OpenAlex using the Aurora Universities SDG Classifier with a cut-off value of 0.4 as a compromise of achieving high recall and precision (OurResearch, 2025). Details about the classification algorithm were provided by Vanderfeesten, Jaworek, and Keßler (2022).

Data

We used an OpenAlex snapshot from August 2024 available to us via the German 'Kompetenznetzwerk Bibliometrie' (Schmidt et al., 2024). We extracted the SDG-relevant publications for (i) USA, (ii) Japan, and (iii) the world in the time period from 2014 to 2023. No restrictions on document types were imposed. Country information was extracted from the author's affiliations. Documents with multiple affiliations were fully counted as a paper for each of the collaborating co-authors. Thus, it is possible that some documents are counted for both countries included in this analysis. Table 1 shows the 17 SDGs with their number of papers in the time period investigated.

Table 1. SDGs with their number of papers in OpenAlex for the time period from 2014 to 2023 ordered decreasingly by the number of papers.

SDG		#Papers	%Papers
3	Good health and well-being	9,603,428	18.85
4	Quality education	5,244,104	10.29
7	Affordable and clean energy	4,326,404	8.49
2	Zero hunger	4,137,054	8.12
10	Reduced inequalities	3,733,772	7.33
16	Peace, justice, and strong institutions	3,634,677	7.13
8	Decent work and economic growth	2,971,404	5.83
11	Sustainable cities and communities	2,692,363	5.28
5	Gender equality	2,427,416	4.76
6	Clean water and sanitation	2,087,444	4.10
14	Life below water	2,030,626	3.99
9	Industry, innovation, and infrastructure	1,951,836	3.83
15	Life on land	1,909,980	3.75
13	Climate action	1,677,011	3.29
17	Partnership for the goals	1,160,840	2.28
1	No poverty	760,432	1.49
12	Responsible consumption and production	604,498	1.19

Overlay maps

Base maps have been used to create overlay maps. Base maps are intended to spatially position concepts from OpenAlex on a map based on citation relations between the concepts. In OpenAlex, concepts are abstract ideas that scholarly works are about. Concepts are assigned to works based on the title, abstract, and the title of the host venue using an automated classifier. Each work is tagged with multiple concepts although some works are not assigned to any concept. We indicated a concept where a country has reached or surpassed 10% of the world-wide SDG-relevant output in a concept with a red dot on the map. Concepts in which a country did not reach this 10% threshold are shown as gray dots. Thus, red dots indicate concepts with many publications of a country that are relevant for the worldwide research targeting SDGs. Data analysis and graphic production have been done using R (R Core Team, 2021) with the R packages 'tidyverse' (Wickham, 2017) and 'ggforce' (Pedersen, 2024).

We used the global base map for OpenAlex (2008-2022) as provided by Haunschild and Bornmann (2024a, 2024b). The base map provides coordinates for the concepts of level 0,1, and 2 of the science covered by OpenAlex. Concepts are one of the field classifications provided by OpenAlex. The maps also include a cluster assignment that is interpreted as a broad scientific classification: (i) Social Sciences and Humanities, (ii) Medicine, (iii) Physics and Engineering, (iv) Mathematics, Computer Sciences, and Theoretical Physics, (v) Biology, and (vi) Chemistry and Material Sciences.

Results

Figure 1 shows the overlay map for the USA. The six different scientific areas are roughly marked with circles and labels.

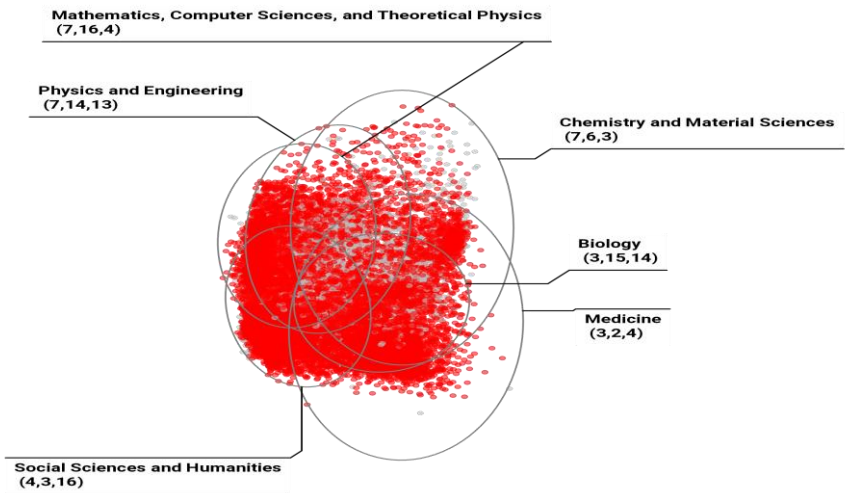


Figure 1. Overlay map of the USA where red dots show concepts with many SDG-relevant publications. The labels of the broad areas are extended by the top 3 SDG numbers in parentheses occurring in these areas (see Table 1).

The overlay map indicates by the many red dots that the USA surpasses the 10% threshold in many concepts (i.e., indicating high SDG-relevance in these fields) within all six broad scientific areas. This is not unexpected due to the very high publication output of the USA in general. The labels of the broad areas were extended by the top 3 SDG numbers in parentheses occurring in these areas (see Table 1). For example, in the case of Physics and Engineering, the top 3 SDGs are 7 ('Affordable and clean energy'), 14 ('Life below water'), and 13 ('Climate Action'). Overall, nine different SDGs occur as top 3 SDGs across all six different broad areas of science.
Figure 2 shows the overlay map for Japan. Due to the lower overall publication output of Japan compared to the USA, fewer red dots are visible. However, several red dots are visible in all broad scientific areas. Overall, eleven different SDGs occur as top 3 SDGs across all six different broad areas of science.
As Figure 2 reveals the map is able to point to Japanese research areas where the country significantly contributed to worldwide SDG relevant research. Two areas of aggregation of red dots indicating high SDG-relevance of Japanese research can be found in the lower-right part (Medicine) and middle-right part (Chemistry and Material Sciences) of the map. In the following, we will have a closer look at these two aggregations of concepts with high SDG-relevance.

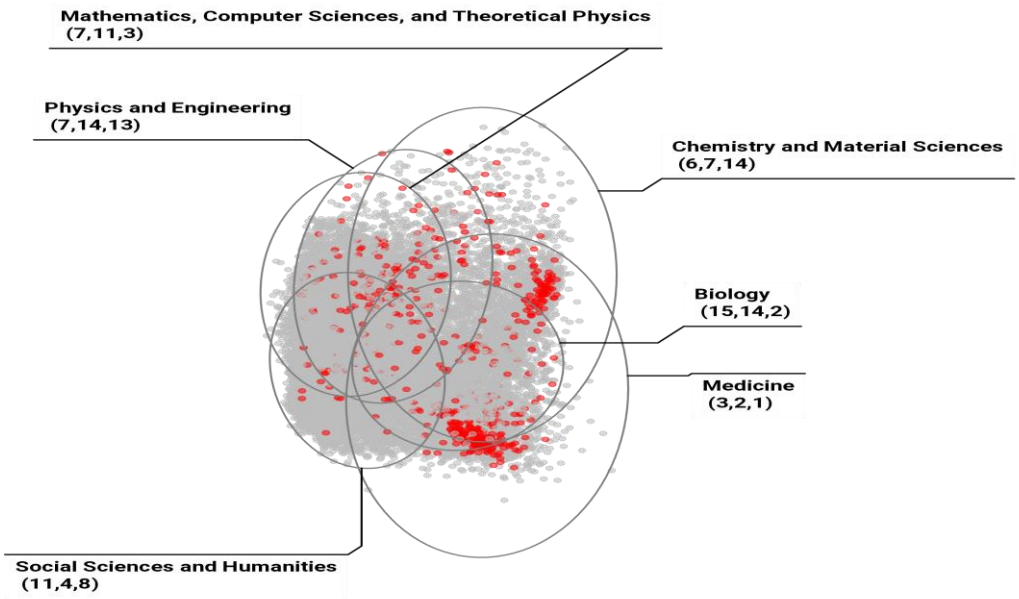


Figure 2. Overlay map of Japan where red dots show concepts with many SDG-relevant publications. The labels of the broad areas are extended by the top 3 SDG numbers in parentheses occurring in these areas (see Table 1).

Concepts of high SDG-relevance in Medicine

Very prominent concepts by the number of SDG-relevant publications within the medical area of concept aggregation with many SDG-relevant publications are the concepts 'Resection', 'Dissection (medical)', and 'Aneurysm'. SDGs 3 ('Good health and well-being'), 2 ('Zero hunger'), and 1 ('No poverty') are the three most relevant SDGs for these concepts for Japanese publications. With the three SDGs, the medical area of concept aggregation reflects the medical research area as a whole. The concept 'Resection' refers to the surgical removal of all or part of an organ, tissue, or biological structure. The concept 'Dissection (medical)' refers to a tear within the wall of a blood vessel. The concept 'Aneurysm' refers to a bulge in the wall of a blood vessel.

Concepts of high SDG-relevance in Chemistry and Material Sciences

Within Chemistry and Material Sciences, concepts such as 'Total Synthesis', 'Diastereomer', and 'Trimethylsilyl' occur with a very high number of SDG-relevant publications in the area of concept aggregation with many SDG-relevant publications. The three concepts exhibit high numbers of publications in SDG 6 ('Clean water and sanitation'). The concept 'Total Synthesis' also contains many publications in SDG 14 ('Life below water'). The concept 'Total Synthesis' refers to a specialized area in organic chemistry that is concerned with synthesizing complex chemical compounds from substances found in nature. The concept 'Diastereomer' describes a specific type of stereoisomer within a compound. This concept also is closely related to organic chemistry. The concept 'Trimethylsilyl' refers to a functional group that is often used as a protective group in certain steps of chemical reactions.

Discussion

Following CoARA that emphases the need for a more holistic approach to research evaluation, we introduce here an approach that is based on overlay maps. The approach reveals national research areas contributing significantly (i.e., more than 10%) to the worldwide SDG-relevant research. We demonstrate our approach using the US and Japanese research as examples. Since the USA is one of the most research active countries in most disciplines, the US map also reveals high research activity with SDG relevance in most disciplines. Our approach is especially interesting for smaller countries with less publications than the USA to reveal their specific contributions to worldwide SDG-relevant research. In this study, we could identify two areas of Japanese research with high relevance for targeting SDGs: Medicine as well as Chemistry and Material Sciences. Since the development of our SDG overlay approach is research in progress, we plan to produce overlay maps also for other countries to reveal their specific SDG-relevant research.

It is one problem of the movements for reforming research assessments that they have not found broad acceptance and application. The results of Rushforth and de Rijcke (2024) show that "there is not yet a deep level of familiarity with international reform movements for responsible metrics and assessment in the United States. The lack of familiarity with the responsible metrics movements' 'responsibility

language' was manifest in: the lack of referencing specific points in responsible metrics statements; lack of awareness of the actors involved in enacting performative powers of metrics (e.g. nobody mentioned publishers); the propensity to present their own 'bottom up' responsibilities which were different from the reform movements' language, or were similar only by coincidence because all actors inhabit the same professional world". The study of Morgan-Thomas, Tsoukas, Dudau, and Gąska (2024) points out that "the limited incidence of non-journal outputs in institutional submissions, the high correspondence between expert score and an aggregate metrics (journal rank), and the non-significance of DORA affiliation, all point to declarations being potentially decoupled from practices". Since our approach is based on freely available OpenAlex data and targets a very relevant question in the area of societal impact measurements, i.e., national contributions to worldwide SDG-relevant research, we assume that there will be a 'market' for its application. We provide an overlay approach that goes beyond using simple counting of publications or citations by displaying national fields with high proportions of SDG-relevant publications on a map and discussing the results for different fields and SDGs.

Our approach is affected by an important limitation that has been addressed, e.g., by Mutz, Bornmann, and Haunschild (2025): the low agreement of different approaches for assigning SDGs to papers. In this study, we used the Aurora Universities SDG Classifier; other classifiers will probably lead to different assignments.

Acknowledgments

Access to OpenAlex bibliometric data has been supported via the German Competence Network for Bibliometrics, funded by the Federal Ministry of Education and Research (grant number: 16WIK2101A

References

Adams, J., McVeigh, M., Pendlebury, D., & Szomszor, M. (2019). *Profiles, not metrics*. Philadelphia, PA, USA: Clarivate Analytics.

Bornmann, L. (2012). Measuring the societal impact of research. *EMBO Reports, 13*(8), 673-676.

Bornmann, L. (2013). What is societal impact of research and how can it be assessed? A literature survey. *Journal of the American Society of Information Science and Technology, 64*(2), 217-233.

Ciarli, T. (Ed.). (2022). *Changing directions: Steering science, technology and innovation towards the sustainable development goals*. SPRU, University of Sussex, UK: STRINGS.

CoARA. (2022). Agreement on Reforming Research Assessment. Retrieved February 7, 2025, from https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf

Haunschild, R., & Bornmann, L. (2024a). Global base maps produced with OpenAlex. Retrieved February 20, 2025, from https://doi.org/10.17617/1.daf7-fq06

Haunschild, R., & Bornmann, L. (2024b). The use of OpenAlex to produce meaningful bibliometric global overlay maps of science on the individual, institutional, and national levels. *PLOS ONE, 19*(12), e0308041. doi: 10.1371/journal.pone.0308041.

Haunschild, R., Bornmann, L., & Adams, J. (2019). R package for producing beamplots as a preferred alternative to the h index when assessing single researchers (based on downloads from Web of Science). *Scientometrics, 120*(2), 925-927. doi: 10.1007/s11192-019-03147-3.

Hernandez, C., Suazo López, F., & Domínguez Pacheco, F. A. (2022). The Sustainable Development Goals index: An analysis (2000-2022). *Transdisciplinary Journal of Engineering & Science, 13*(0). doi: 10.22545/2022/00213.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429-431.

Morgan-Thomas, A., Tsoukas, S., Dudau, A., & Gąska, P. (2024). Beyond declarations: Metrics, rankings and responsible assessment. *Research Policy, 53*(10), 105093. doi: 10.1016/j.respol.2024.105093.

Mutz, R., Bornmann, L., & Haunschild, R. (2025). How to use assignments of United Nations sustainable development goals (SDGs) to scientific papers in research evaluation? The proposal of a gold standard combining assignments from different data providers. *Scientometrics*. doi: 10.1007/s11192-025-05254-w.

OurResearch. (2025). How do you classify works as contributing to the UN SDGs? Retrieved February 24, 2025, from https://help.openalex.org/hc/en-us/articles/27972124390679-How-do-you-classify-works-as-contributing-to-the-UN-SDGs

Pedersen, T. L. (2024). ggforce: Accelerating 'ggplot2'. Retrieved February 24, 2025, from https://CRAN.R-project.org/package=ggforce

Purnell, P. J. (2022). A comparison of different methods of identifying publications related to the United Nations sustainable development goals: Case study of SDG 13: Climate action. *Quantitative Science Studies, 3*(4), 976-1002. doi: 10.1162/qss_a_00215.

R Core Team. (2021). R: A language and environment for statistical computing. Retrieved February 20, 2025, from https://www.R-project.org/

Rushforth, A., & de Rijcke, S. (2024). Practicing responsible research assessment: Qualitative study of faculty hiring, promotion, and tenure assessments in the United States. *Research Evaluation, 33*, rvae007. doi: 10.1093/reseval/rvae007.

Rushforth, A., & Hammarfelt, B. (2023). The rise of responsible metrics as a professional reform movement: A collective action frames account. *Quantitative Science Studies, 4*(4), 879-897.

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., . . . Stahlschmidt, S. (2024). The data infrastructure of the German Kompetenznetzwerk Bibliometrie: An enabling intermediary between raw data and analysis. Retrieved October 28, 2024, from https://doi.org/10.5281/zenodo.13935407

Thelwall, M. (2024). Quantitative methods in research evaluation: Citation indicators, altmetrics, and artificial intelligence. Retrieved July 5, 2024, from https://arxiv.org/abs/2407.00135

Vanderfeesten, M., Jaworek, R., & Keßler, L. (2022). AI for mapping multi-lingual academic papers to the United Nations' Sustainable Development Goals (SDGs). Retrieved February 20, 2025, from https://dx.doi.org/10.5281/zenodo.5603019

Wickham, H. (2017). Tidyverse: Easily install and load the 'Tidyverse'. R package version 1.2.1. Retrieved June 22, 2020, from https://CRAN.R-project.org/package=tidyverse

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., . . . Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. Bristol, UK: Higher Education Funding Council for England (HEFCE).

https://doi.org/10.51408/issi2025_158

Does Evaluating Research Still Need Virtues in the Age of ChatGPT?

Alessio Vaccari[1], Cinzia Daraio[2]

[1]*alessio.vaccari@uniroma1.it*
Department of Philosophy, Sapienza University of Rome, Rome (Italy)

[2]*daraio@diag.uniroma1.it*
Department of Computer, Control and Management Engineering "Antonio Ruberti" (DIAG), Sapienza University of Rome, Rome (Italy)

Abstract

In the era of Artificial Intelligence (AI)-driven research, the evaluation of scientific work must go beyond the assessment of results and consider the intellectual virtues of researchers. This article explores the role of intellectual virtues - such as open-mindedness, courage and conscientiousness - in ensuring ethical and epistemically sound research. Drawing on key philosophical perspectives, including those of Sosa, Zagzebski, and Pritchard, we argue that intellectual virtues remain essential even as AI tools, such as ChatGPT, reshape cognitive processes. While AI may reduce reliance on internal cognitive skills, it need not diminish intellectual virtues; rather, these virtues guide the responsible and reflective use of AI in research. We also propose a virtue-based framework for research evaluation that distinguishes between different researcher archetypes and emphasises the role of practical wisdom (*phronesis*) in dealing with ethical dilemmas. Ultimately, we argue that research evaluation in the AI era must prioritise intellectual virtues in order to maintain integrity, foster innovation, and ensure that AI tools serve as supportive tools rather than replacing human intellectual effort.

Introduction

The increasing integration of Artificial Intelligence (AI) in research practices is reshaping the landscape of academic inquiry, challenging traditional paradigms of knowledge production, evaluation, and intellectual engagement. AI-powered tools like ChatGPT have demonstrated their capacity to assist researchers in a variety of tasks, from literature review and data synthesis to writing and argumentation. While these advancements hold the potential to accelerate research processes and enhance accessibility, they also raise significant epistemological and ethical concerns. One pressing issue is whether the reliance on AI in academic work risks undermining the intellectual virtues that have historically underpinned rigorous and ethical research.

Intellectual virtues—such as open-mindedness, intellectual courage, conscientiousness, and epistemic humility—have long been regarded as essential qualities of good scholarship. These virtues guide researchers in critically evaluating evidence, engaging with diverse perspectives, and exercising sound judgment in the pursuit of knowledge. However, as AI increasingly automates cognitive tasks, there is a growing concern that it may foster intellectual passivity, reducing the researcher's engagement in deep, reflective thinking. This raises fundamental questions: Can intellectual virtues survive in an AI-dominated research

environment? How should research evaluation adapt to ensure that AI tools support rather than replace human intellectual effort?

This paper explores these questions through the lens of virtue epistemology, drawing on the philosophical perspectives of thinkers such as Ernest Sosa, Linda Zagzebski, and Duncan Pritchard. These scholars argue that intellectual virtues are not merely instrumental to knowledge acquisition but are constitutive of a well-functioning intellectual character. Their insights provide a valuable framework for understanding how researchers can engage with AI in ways that preserve and even enhance intellectual virtues, rather than allowing technology to erode them. We argue that while AI can alter cognitive processes by reducing reliance on certain internal skills, it does not inherently threaten intellectual virtues. Instead, the responsible and reflective use of AI—guided by virtues—can ensure that these technologies serve as powerful tools for knowledge advancement rather than as substitutes for human intellectual effort.

To address these concerns, we propose a virtue-based framework for research evaluation that extends beyond traditional metrics of output assessment. This framework distinguishes between different researcher archetypes—such as the Good Researcher, who exemplifies intellectual virtues in creative knowledge advancement; the Leader Researcher, who combines intellectual and social virtues to inspire ethical research practices; and the Honest Researcher, who upholds integrity and reliability, often at the early stages of their academic career. By incorporating intellectual virtues into research evaluation, we advocate for an approach that prioritizes not only the validity and impact of research but also the ethical and epistemic character of those who produce it.

By engaging with these themes, we seek to contribute to the ongoing debate on the ethical and epistemic challenges of AI in research. We argue that, rather than diminishing intellectual virtues, AI should be integrated into academic practices in a way that fosters critical engagement, intellectual responsibility, and ethical integrity.

The paper is structured as follows. Section 2 examines the role of intellectual virtues in research, outlining key philosophical perspectives on virtue epistemology and their relevance to academic inquiry. Section 3 explores the challenges posed by AI technologies in research practices, particularly the potential risks of cognitive diminishment and ethical dilemmas in AI-assisted scholarship. Section 4 introduces a virtue-based framework for research evaluation, distinguishing between different researcher archetypes and emphasizing the importance of practical wisdom (*phronesis*) in navigating ethical challenges. Section 5 discusses the implications of AI in research evaluation, considering how AI tools can support the exercise of intellectual virtues rather than undermine them. Section 6 concludes the paper by reaffirming the necessity of intellectual virtues in research evaluation and proposing directions for future research on the ethical integration of AI in academia.

Intellectual Virtues and Research Evaluation

The evaluation of research practices should extend beyond assessing research outputs alone. It must also consider the moral and intellectual character of researchers, who play a crucial role in the research process. Intellectual virtues—such as courage, open-mindedness, and conscientiousness—are essential for

advancing knowledge and maintaining ethical research standards. Integrating these virtues into the evaluation framework contributes to a more comprehensive and meaningful assessment, i.e., a 'good' evaluation, of research practices (Daraio & Vaccari, 2020; 2022).

Challenges Posed by AI Technologies

The rise of generative AI tools like ChatGPT introduces potential risks of cognitive diminishment, where overreliance on technology undermines critical cognitive abilities. This raises pressing questions: Can intellectual virtues survive in an age dominated by AI? And how can research evaluation systems adapt to ensure these virtues remain central?

To address these challenges, we propose to use the theory of intellectual virtues as articulated by thinkers like Ernest Sosa, Linda Zagzebski, and Duncan Pritchard. These theories emphasize that intellectual virtues are not merely instrumental but are constitutive of the good human life, offering a pathway to deeper understanding rather than just factual knowledge (Pritchard 2015, Zagzebski 1996, Sosa 1980).

Key Philosophical Perspectives on Intellectual Virtues and Their Role in Research Evaluation

Three distinct perspectives on intellectual virtues merit examination. In this section, we will explore the first two, while the third will be discussed separately. The first influential model is that developed by Ernest Sosa (Sosa, 1980, 1981, 1985; Greco, 2002). According to Sosa, intellectual virtues are innate or acquired dispositions that reliably lead to grasping truth and avoiding falsehood. He used this concept to develop a theory of epistemic justification that overcomes the challenges posed by foundationalism and coherentism. In his model

- A belief B(p) is epistemically justified for a person S (justified in the sense required for knowledge) if and only if B(p) is produced by one or more intellectual virtues of S (Sosa, 1985, p. 290).

Epistemic principles become dispositions to form true beliefs about the environment on the basis of sensory inputs of different modalities. Because these powers and capacities are reliable (memory, introspection, logical intuition), they give rise to epistemic justifications for their respective products.

Similarly, he argues that various kinds of deductive or inductive reasoning - together with coherence-seeking reason - are virtuous because they reliably lead one from true belief to further true belief.

A second line of research has instead identified intellectual virtues with personality traits or qualities of character. According to Montmarquet, the intellectual virtues - such as intellectual courage and intellectual prudence - are analogous to the moral virtues (such as moral temperance and moral courage) in at least two ways:

1. The intellectual virtues have a passionate and motivational component, they are constitutively linked to the desire for truth (Montmarquet, 1993).

2. The exercise of the intellectual virtues is under our control: although we cannot control our perceptual impressions, we can control whether or not we take an idea seriously or whether or not we choose to consider a line of argument accurately (Montmarquet, 1993).
3. Intellectual virtues, like moral virtues, are appropriate objects of praise and blame (Montmarquet, 1993).

Along the same theoretical line is the position of Zagzebski, who, more than Montmarquet, emphasised the closeness of the moral and intellectual virtues. Like the moral virtues, the intellectual virtues involve a general motivation to achieve true belief and are reliably successful in doing so. But because the true is a component of the good, Zagzebski argues, the intellectual virtues can be understood as a subset of the moral virtues.
According to Zagzebski, an advantage of understanding intellectual virtue in this way is that it allows for an understanding of knowledge. She argues that:

- An act of intellectual virtue A is an act that arises from the motivational component of A, is something that a person with virtue A would (probably) do in the circumstances, is successful in achieving the end of A's motivation, and is such that the agent acquires a true belief through these features of the act (Zagzebski, 1996, p. 270).

For Zagzebski, an advantage of understanding intellectual virtue in this way is that it allows an understanding of the knowledge. More precisely: S has knowledge of P if

1. p is true, and
2. The true belief B (p) of p arises from the acts of an intellectual virtue.

Therefore, S has knowledge of p if belief p arises from actions of intellectual virtues (Zagzebski, 1996, pp. 264-3).
Having outlined - albeit schematically - the main positions on the nature of the intellectual virtues, let us make some general points on the nature of intellectual virtues:

1. Despite the differences between these two models, none seems to explicitly identify the virtues with cognitive abilities, understood as something that is clearly distinct from the motivational components of virtue.
2. The desiderative components seem to be constitutive elements of the intellectual virtues. Although Sosa does not explicitly include them, it seems implausible not to include something of the sort in his characterisation of the search for consistency between perceptions.

Intellectual Virtues and Cognitive Abilities. Pritchard's Model

Based on these general considerations, we believe that the conception of the intellectual virtues as recently articulated by Duncan Pritchard captures the essential elements of these virtues. Pritchard is one of the most authoritative proponents of the

so-called virtue responsibility conception, which places the cognitive character of the agent at the centre of his analysis. He claims:

"Virtue epistemology puts the cognitive character of the subject centre-stage, where this means the interconnected web of the subject's integrated cognitive faculties, cognitive abilities and intellectual virtues" (Pritchard, 2015, p. 3; see also Axtell, 1997; Kvanvig, 2010, and Greco, 2011).

According to Pritchard, the cognitive character of the subject is not reducible to virtues, but is identified with "an integrated network of cognitive skills, cognitive abilities and intellectual virtues". Let us distinguish these elements and see how they relate to each other.

1. *Cognitive faculties*: these are the innate cognitive abilities that individuals possess, such as those involved in perception or memory. They can be improved through training, which usually involves integrating the faculty with other cognitive traits.
2. *Cognitive abilities*, on the other hand, are acquired rather than innate and involve specific skills - such as the facility to do arithmetic. Acquired cognitive skills draw on existing cognitive abilities and are used to perform specific cognitive tasks.
3. *Intellectual virtues*: Although they are similar to cognitive skills in that they are acquired cognitive traits that draw on innate cognitive faculties, they differ significantly from them. For example, the exercise of an intellectual virtue not only facilitates access to truths, but also manifests the subject's motivation to acquire truth. Similar to Montmarchet and Zag, intellectual virtues express our love of truth (Pritchard, 2016; see also Zagzebski, 1996). Cognitive virtues are typically not accompanied by such a motivational component, but rather are associated with the desire to be better at a particular task than a competitor.

Pritchard highlights two important distinctions between cognitive abilities and intellectual virtues:

A. Intellectual Virtues - like moral virtues - are constitutive elements of the good human life. They therefore possess a special axiological status that cognitive abilities do not. The latter have only an instrumental value. Virtues, on the contrary, have value for those who possess them, regardless of their «practical usefulness» (Pritchard, 2014, p. 4). Intellectual virtues thus have value for themselves as manifestations of cognitive agency (Pritchard, 2014, p. 4; Roberts & Wood, 2007).

> <<… while the wise person would not willingly give up an intellectual virtue, he might choose to give up a cognitive skill if it ceased to be practically useful>> (Pritchard & Turri, 2011; see also Pritchard, 2007).

B. A further axis of differentiation is in terms of specificity. Cognitive skills tend to be understood in a narrow sense, in the sense that they are often abilities to reliably perform specific cognitive tasks (e.g. simple arithmetic). Intellectual virtues, on the other hand, are very broad cognitive traits of the

agent, such as conscientiousness, open-mindedness, etc.. This reflects the general regulative function that intellectual virtues tend to play within a subject's cognitive economy, in that they guide the employment of one's cognitive abilities and faculties, rather than vice versa.

Addressing the Challenges of AI Technologies: Implications for Research Evaluation

Through these lenses, we propose that the decline in cognitive abilities from AI use does not necessarily erode intellectual virtues. Instead, these tools can complement virtues by facilitating reflective and critical engagement with AI outputs (Cassinadri, 2024).

More precisely, while it is true that the use of ChatGPT and other generative AI tools can have the effect of weakening internal cognitive abilities, this does not necessarily have a negative impact on intellectual virtues:

1. Although virtues and cognitive capacities cooperate with each other in the construction of true representations of the world - and in this sense they are concomitant factors. They are different psychological factors. Summarising Pritchard's lesson: whereas the function of cognitive abilities is to enable the acquisition of a set of true factual information (Cassinadri, 2024, p. 4), the function of virtues is to acquire 'understanding' (Cassinadri, 2024, p. 4; Pricthard, 2013, 2016; Mollick & Mollick, 2022).
2. In contrast to the mere possession of true beliefs, 'undestanding' denotes the knowledge that the agent possesses when (a) he is aware that the sources of his beliefs are reliable and (b) he knows the reasons why this is so. In this way, the virtuous subject is a cognitive agent and not merely a subject who holds true beliefs.
3. Although the use of ChatGPT could in principle lead to cognitive diminishment due to the fact that we overuse technology at the expense of exercising cognitive skills, this may not be as disastrous an outcome as it seems. After all, once the outputs from these technological tools are screened by the intellectual virtues, these outputs can become a potentially useful source of information to be evaluated reflectively and critically like any other cognitive output.
4. The development of intellectual virtue need not depend on the outputs of pure cognitive abilities, but may also derive from the outputs of AI-supported technologies. In both case, they are a starting point for the understanding of reality made possible by the intellectual virtues.

If intellectual virtues remain intact through the use of AI, then concerns about AI-induced cognitive decline may be less troubling than they appear. This has significant implications for how we assess the quality and integrity of research practices in an AI-driven landscape.

The proper use of AI tools and the intellectual virtues

Having shown how intellectual virtues are not undermined by the use of AI, it is possible to argue that the proper use of these tools requires the possession of applying intellectual virtues.

In doing so, we intend to extend Kristjánsson and Fowers' approach (Kristjánsson and Fowers, 2024), in particular their exploration of *phronesis* (practical wisdom) in professional ethics, to ethical considerations of AI tools in research.

Kristjánsson and Fowers' approach emphasises the importance of cultivating intellectual virtues in professional ethics, particularly when navigating complex and morally charged situations. They argue that *phronesis* should guide professionals in making ethical decisions, especially in situations where shared rules may not suffice. In this context, we can apply this framework to the ethical use of AI tools such as ChatGPT in research.

Intellectual virtues - such as open-mindedness, intellectual courage, intellectual humility and intellectual perseverance- can offer a lens through which to evaluate the use of AI tools in academic practices. When using ChatGPT for research, these virtues help to ensure that AI tools are exploited ethically and improve the overall quality of research, rather than reducing it.

For instance, researchers may need open-mindedness, being receptive to the new knowledge that AI tools can provide, without relying on them as the sole source of information. Furthermore, they should possess intellectual perseverance, continuing to rigorously evaluate, cross-reference and verify AI-assisted results in the research process, ensuring that AI does not merely simplify tasks, but instead contributes significantly to the discovery and understanding of knowledge. Another fundamental virtue is transparency, which requires researchers to clearly disclose how AI tools were used in their research process (methodology, data analysis, etc.).

Research Practices, Intellectual Virtues and AI

The crucial role that virtues play in the correct use of AI tools becomes even more apparent if we address the question of what constitutes a good evaluation of research itself. Again, we can extend Kristjánsson and Fowers' phronesis-focused framework to emphasise the evaluation of the research practices they use: how AI tools help to evaluate the practices behind the research, not just the research results themselves. This is particularly congenial to our approach to evaluating scientific research. Building on the theoretical foundations of intellectual virtues, we have characterized academic/scientific research as a socially established cooperative human activity (Daraio and Vaccari 2020). Following MacIntyre, we define a good social practice as

> - "[…] any coherent and complex form of socially established cooperative human activity through which goods internal to that form of activity are realized in the course of trying to achieve those standards of excellence which are appropriate to, and partially definitive of, that form of activity, with the result that human powers

to achieve excellence, and human conceptions of the ends and goods involved, are systematically extended" (MacIntyre, 1981 first ed.; pp. 1985, 187).

On the basis of the definition of good social practice, we characterize a good research practice as

- "[…] any coherent and complex form of socially established cooperative human activity through which its participants, through the exercise of a set of refined human psychological qualities or virtues, contribute to the advancement of the body of knowledge that is constitutive of that practice in a way that has a positive impact on the lives of researchers and society as a whole" (Daraio and Vaccari 2020, p. 1059).

Good evaluation of research practices must use a holistic approach to the evaluation of research practices that examines methodological soundness, ethical rigour and validity of conclusions. AI can support this by offering quick access to related literature, providing computational assistance for data analysis and highlighting potential flaws or inconsistencies. It can be argued that the overall quality of this evaluation depends on the way AI tools are integrated and the intellectual virtues applied to their use.
Evaluators should use AI not only to assess individual research projects, but also to reflect on broader trends in research practices, such as the use of AI itself.
This includes examining the impact of AI on ethical decision-making, research design and data processing. Intellectual virtues such as intellectual courage can help evaluators ask difficult questions about the ethical use of AI tools in the research process.
Furthermore, the integration of AI tools such as ChatGPT into research evaluation can improve the process by providing computational assistance and expanding access to information. However, the quality of the evaluation of research practices depends significantly on how these tools are used. By applying intellectual virtues such as open-mindedness, intellectual humility, integrity, accountability and critical thinking, researchers and evaluators can ensure that AI tools support, rather than undermine, the ethical rigor and methodological soundness of research evaluation.

Using virtues in AI: Three Types of Researchers

Building on the three types of researchers outlined in our previous work (Daraio & Vaccari, 2020; 2022), we apply them to the challenges of integrating artificial intelligence into research practices:

1. *The Leader Researcher*: This role combines intellectual and social virtues to inspire excellence and collaboration. The Leader sets ethical standards for the use of AI within their teams and, together with the Good Researcher, embodies virtues such as conscientiousness and open-mindedness. They ensure that AI tools like ChatGPT are integrated in ways that align with both the ethical and epistemic goals of the research teams.

2. *The Good Researcher*: A model of intellectual virtues, the Good Researcher advances knowledge creatively while adhering to ethical and epistemic standards. Alongside the Leader, they embody virtues such as conscientiousness and open-mindedness, ensuring that AI tools like ChatGPT complement—not replace—the intellectual effort. They maintain a reflective and critical engagement with AI outputs, ensuring these tools align with the broader goals set by the Leader.
3. *The Honest Researcher*: Committed to upholding ethical standards, the Honest Researcher is a reliable contributor, typically early in their career. They assist the Leader and Good Researcher in applying these principles, learning from their guidance and experience.

These roles illustrate how intellectual virtues translate into tangible contributions to research practices. However, the integration of AI in research raises important ethical dilemmas. For example, does reliance on tools like ChatGPT undermine intellectual rigor, or can it enhance inclusivity and creativity? Tools for detecting AI-generated content underscore the increasing need for ethical guidelines in research practices (Mateos-Sanchez et al., 2022).

Virtuous researchers navigate these dilemmas by critically evaluating AI-generated outputs and ensuring their use aligns with the pursuit of deeper understanding, rather than simply serving utility-driven goals.

Conclusion

Intellectual virtues enable researchers to make ethical and effective use of tools such as ChatGPT, thereby fostering understanding and innovation. By aligning theoretical insights with practical applications, we can ensure that research practices continue to meet the highest standards of excellence and integrity.

As AI becomes increasingly embedded in research practices, it is imperative to reassess the criteria by which scholarly work is evaluated. This paper has argued that research evaluation must extend beyond output-based metrics to consider the intellectual virtues that shape ethical and epistemically responsible inquiry. Intellectual virtues—such as open-mindedness, intellectual courage, conscientiousness, and epistemic humility—are not only fundamental to sound research but also serve as safeguards against the risks posed by the growing reliance on AI in academic work.

Through an engagement with virtue epistemology, particularly the perspectives of Ernest Sosa, Linda Zagzebski, and Duncan Pritchard, we have highlighted the distinction between cognitive abilities and intellectual virtues. While AI can enhance cognitive abilities by providing rapid access to information, generating text, and automating certain tasks, it does not cultivate intellectual virtues on its own. Instead, the responsible and reflective use of AI requires researchers to exercise virtues that ensure AI tools support, rather than replace, human intellectual effort. The ethical integration of AI in research thus depends on fostering a culture of intellectual virtue, where researchers remain actively engaged in critical thinking, methodological rigor, and ethical accountability.

A key contribution of this paper is the virtue-based framework for research evaluation, which proposes a holistic approach to assessing research. By distinguishing between different researcher archetypes—the Good Researcher, the Leader Researcher, and the Honest Researcher—we have emphasized that scholarly excellence is not solely determined by knowledge production but also by the intellectual character and ethical integrity of researchers. These archetypes illustrate how intellectual virtues manifest in academic work, shaping both individual research practices and the broader research community. Moreover, the concept of practical wisdom (*phronesis*) has been introduced as a guiding principle for navigating the ethical dilemmas posed by AI in academic settings.

In response to the question posed in the title—*Does evaluating research still need virtues in the age of ChatGPT?*—our answer is a clear and affirmative yes. Even though AI can assist in cognitive tasks and streamline the research process, the evaluation of research still requires human judgment guided by intellectual virtues. These virtues ensure that the use of AI remains critical, ethically aware, and epistemically responsible, thereby safeguarding the integrity and meaningfulness of academic work.

Beyond its theoretical contributions, this paper also raises critical questions about the future of AI-assisted research. As AI continues to advance, it is likely to play an even more significant role in shaping academic inquiry. This evolution presents both opportunities and challenges. On one hand, AI has the potential to democratize access to knowledge, reduce cognitive load, and facilitate interdisciplinary collaboration. On the other hand, the overreliance on AI could lead to intellectual complacency, where researchers passively accept AI-generated outputs without critical engagement. Ensuring that AI remains a tool for augmentation rather than replacement requires active reflection on the principles that govern its use.

The practical implications of our argument suggest that research institutions, funding bodies, and academic journals should revise their evaluation criteria to include the demonstration of intellectual virtues. This might include explicit guidelines for ethical AI use, reflective commentary on methodological choices, or assessments of epistemic responsibility.

Given the profound impact of AI on research practices, future studies should further investigate the following aspects.

While this paper has provided a theoretical foundation, empirical research is needed to assess whether AI affects researchers' intellectual virtues in practice. Studies could explore whether frequent reliance on AI tools correlates with changes in researchers' critical thinking skills, epistemic humility, or intellectual perseverance.

As AI becomes increasingly integrated into research methodologies, academic institutions and funding bodies should consider incorporating virtue-based principles into research evaluation criteria. Future research could contribute by formulating guidelines on how intellectual virtues should be assessed in AI-assisted research environments.

Beyond theoretical discussions, it is essential to explore concrete strategies for fostering intellectual virtues among researchers who engage with AI. Educational programs, mentorship models, and institutional policies could be designed to

encourage the cultivation of virtues such as open-mindedness, conscientiousness, and intellectual humility.

While AI can enhance research productivity, it also introduces ethical dilemmas regarding authorship, plagiarism, and the reliability of AI-generated content. Further exploration is needed to develop mechanisms that ensure transparency, accountability, and fairness in AI-assisted research.

In sum, we conclude that evaluating research still unequivocally requires intellectual virtues—even, and especially, in the age of ChatGPT. By embedding these virtues into research evaluation, we uphold not only the epistemic but also the moral foundations of academic inquiry.

References

Axtell, J. (1997). Recent Work in Virtue Epistemology. *American Philosophical Quarterly*, 34(1), 1–26.

Greco, J. (2002). Virtues and Vices of Virtue Epistemology. *Canadian Journal of Philosophy*, 32(3), 279–300.

Greco, J. (2011). Epistemic Goodness and the Value of Knowledge. In J. Greco (Ed.), *Achieving Knowledge: A Virtue-Theoretic Account of Epistemic Normativity* (pp. 3–21). Cambridge University Press.

Daraio, C., & Vaccari, A. (2020). Using normative ethics for building a good evaluation of research practices: towards the assessment of researcher's virtues. *Scientometrics*, 125(2), 1053-1075.

Daraio, C., & Vaccari, A. (2022). How should evaluation be? Is a good evaluation of research also just? Towards the implementation of good evaluation. *Scientometrics*», 1, pp. 1-20.

Kristjánsson, K., & Fowers, B. J. (2024). Phronesis: *Retrieving Practical Wisdom in Psychology, Philosophy, and Education*. Oxford University Press.

Kvanvig, J. (2010). *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press.

Mateos-Sanchez, R., Gonzalez-Sanchez, M., & Medina, M. (2022). The Influence of Virtue Epistemology on Ethical Decision-Making in Health Professionals. *Journal of Medical Ethics*, 48(10), 707-714.

MacIntyre, A. (1985). *After Virtue: A Study in Moral Theory*. 2nd ed. University of Notre Dame Press

Mollick, E., & Mollick, S. (2022). The Role of Intellectual Virtue in Entrepreneurial Decision-Making. *Journal of Business Ethics*, 181(3), 603–621.

Montmarquet, J. A. (1993). Epistemic *Virtue and Doxastic Responsibility*. Rowman & Littlefield

Nyholm, S. (2023). Artificial Intelligence and the Ethics of Intellectual Virtue. *Journal of Ethics and Information Technology*, 25(1), 45-58.

Pritchard, D. (2007). *The Nature and Value of Knowledge: Three Investigations*. Oxford University Press.

Pritchard, D., & Turri, J. (2011). Epistemic Virtue and the Justification of Belief. In *Virtue Epistemology: Essays on Epistemic Virtue and Responsibility* (pp. 167–184). Oxford University Press

Pritchard, D. (2013). *What Is This Thing Called Knowledge*? 3rd ed. Routledge.

Pritchard, D. (2014). Epistemic Disagreement: Theoretical and Practical Issues. In R. A. Briggs (Ed.), *The Epistemology of Disagreement* (pp. 1-20). Routledge.

Pritchard, D. (2015). *Epistemic Angst: Radical Skepticism and the Groundlessness of Our Believing*. Princeton University Press.
Pritchard, D. (2016). *Epistemic Luck*. Oxford University Press.
Roberts, R. C., & Wood, W. J. (2007). *Intellectual Virtues: An Essay in Regulative Epistemology*. Oxford University Press.
Sosa, E. (1980). The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge. *Midwest Studies in Philosophy*, 5(1), 3–26.
Sosa, E. (1981). The Analysis of Knowledge and the Problem of Luck. In P. A. French, T. E. Uehling Jr., & H. K. Wettstein (Eds.), *Midwest Studies in Philosophy*, 5, 177–192.
Sosa, E. (1985). Knowledge and Intellectual Virtue. *The Monist*, 68(2), 226–245.
Vallor, S. (2015). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. Oxford University Press.
Zagzebski, L. (1996). *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge University Press.

https://doi.org/10.51408/issi2025_139

Towards a Responsible Research Assessment Transition: A Novel Framework for Researcher Profiles

Zenia Xenou[1], Giulia Malaguarnera[2], Lottie Provost[3], Natalia Manola[4]

[1]*zenia.xenou@openaire.eu*, [2]*giulia.malaguarnera@openaire.eu*
OpenAIRE AMKE, Artemidos 6 & Epidavrou 15125 Maroussi, Greece (Athens)

[3]*lottiemiaprovost@cnr.it*
Consiglio Nazionale delle Ricerche, Via Giuseppe Moruzzi, 56124 Pisa (Italy)

[4]*natalia.manola@openaire.eu*
OpenAIRE AMKE, Artemidos 6 & Epidavrou 15125 Maroussi, Greece (Athens)

Abstract

The reform of the research assessment system is a top priority on the European Research Area policy agenda. Recognizing that the evaluation of research projects, researchers, research units, and institutions plays a crucial role in the functioning of a robust Research and Innovation system, recent policy efforts emphasize the need for transformative approaches to research assessment. While research is increasingly collaborative and interdisciplinary, traditional research assessment methods, predominantly reliant on publication metrics, only capture a narrow perspective of the diverse activities that constitute high quality and impactful research.
In response, the CoARA Agreement on Reforming Research Assessment, supported by over 800 signatories, calls for a broader recognition of the breadth and diversity of research contributions, career paths, and outputs. This reform champions an evaluation paradigm that prioritises qualitative assessment, supported by a responsible use of quantitative indicators. The movement for reform also calls for acknowledgement of contributions to Open Science, focusing attention on the need to shift towards more inclusive and transparent evaluation frameworks, supported by open research information and non-proprietary data sources.
This practice-oriented contribution focuses on the development of a framework for Researcher Profile within the Horizon Europe project GraspOS. The Researcher Profile is a service aiming to support research funding and performing organizations in the implementation of CoARA Agreement commitments, and to offer a flexible framework for assessing researchers which values diverse practices and prioritizes comprehensive quality and societal impact of research.

Introduction

In recent years, the European Research Area policy agenda has placed the reform of the research assessment system at the forefront of its policy actions, recognizing that the way research projects, researchers, research units, and research institutions are assessed is fundamental for a well-functioning Research and Innovation system. These policy efforts aim to accelerate the shift away from traditional, publication-based assessment methods, underlining their limitations in reflecting the increasingly collaborative and interdisciplinary nature of research (European Research Area policy agenda, 2022). Consolidated evidence shows that publication-based metrics such as the Journal Impact Factor and the h-index (Elsevier Language Services, 2020) fail to reflect the broad range of activities that make up research and are widely (mis)used as proxies for assessing the quality, performance and impact of research

and researchers (Hicks et al., 2015; Pontika et al., 2022; DORA, 2024). Critics argue that the current assessment system creates perverse incentives, encouraging researchers to prioritize publication venue and citation counts (Edwards et al., 2017) over research quality, open collaboration, and societal impact (Di Donato, 2024).

In response to the identified challenges, the European Commission has led efforts to establish a clear and common direction for the reform of research assessment practices. The European Commission Scoping Report "Towards a reform of the research assessment system" (European Commission, 2021) called for research proposals, researchers, research units and research institutions to be "evaluated on their intrinsic merits and performance rather than on the number of publications and where they are published, promoting qualitative judgement with peer-review, supported by responsible use of quantitative indicators." Echoing this call, signatories of the Agreement on Reforming Research Assessment (ARRA) (CoARA, 2022) have undertaken to uphold a series of commitments, including recognizing and valuing diverse contributions to and careers in research.

A number of EU-funded projects are tasked with supporting the ongoing policy reforms and designing new ways to incentivize higher quality research, collaboration and Open Science practices (European Commision, 2024). Among these, the Horizon Europe project GraspOS: Next Generation Research Assessment to Promote Open Science (GraspOS) addresses the need for new services and tools to support a research assessment system that incentivizes Open Science practices. The project aims to develop data infrastructure facilitating qualitative and quantitative assessments, ultimately supporting the practical implementation of the reform at various levels and the transition towards an Open Science-aware responsible research assessment.

Research Questions: Advancing Fair and Inclusive Research Assessment

Building upon these ongoing reform efforts, this paper introduces an innovative service, the Researcher Profile, designed to promote fair and responsible research assessment. The central research question guiding this development of the service is how responsible and fair research assessment can be effectively promoted while also enabling researchers to showcase their contributions beyond publications. This ongoing research aims to explore alternative and holistic frameworks that recognize diverse research outputs, such as datasets, software, policy impact, public engagement, and interdisciplinary collaborations. By addressing systemic biases in current assessment models, the study seeks to develop equitable and transparent approaches that align with open science principles and foster a more inclusive research environment.

A key focus is on designing a researcher profile service that effectively integrates qualitative descriptions with responsibly used quantitative indicators to enhance research assessment. This involves examining how the developing service contributes to a more inclusive and responsible evaluation of diverse scholarly contributions. By ensuring a balanced approach, the study aims to move beyond traditional publication-centric metrics, enabling a more comprehensive representation of researchers' work. This study will identify key barriers and provide

support to researchers across disciplines in demonstrating the broader impact of their work. Ultimately, it aims to contribute to a more responsible and comprehensive research evaluation system that values diverse scholarly outputs while ensuring fairness and transparency.

The Framework for Researcher Profile: an innovative service to support organizations in adopting responsible research assessment practices

The need for innovative services to support responsible research assessment transition

Our approach integrates the responsible use of quantitative indicators with qualitative information on researchers' contributions, ensuring a more holistic evaluation process. This innovation addresses the need to move beyond purely metric-based assessments by incorporating contextually rich, qualitative insights into research impact, collaboration, and Open Science engagement. Furthermore, our service introduces customizable templates tailored to different research domains. By accounting for the unique requirements and evaluation criteria of various disciplines, these templates enable a more nuanced and equitable representation of researchers' contributions. This customization not only enhances the visibility of diverse research outputs but also facilitates more efficient and meaningful assessments aligned with field-specific expectations.

The Researcher Profile represents a significant step towards responsible research assessment, fostering a system that values quality, collaboration, and societal impact over traditional publication metrics. By aligning with the broader objectives of the ERA policy agenda and recent EU initiatives, our contribution seeks to advance the practical implementation of more inclusive and comprehensive research evaluation practices.This service is envisaged as a service showing a novel researchers' curricula considering a framework of indicators and research activities library. The Researcher Profile design aligns closely with the latest policy recommendations and guidelines promoting a responsible approach to research assessment. In particular, the concept considers the SCOPE Framework (INORMS Research Evaluation Group, 2023) and follows the DORA Guidance on the responsible use of quantitative indicators in research assessment (DORA, 2024).

Leveraging an Innovative Service to Tackle Research Assessment Challenges

Metrics and indicators serve as useful benchmarks for measuring research activities; however, they cannot fully capture the complexities of academic contributions on their own. As outlined in the DORA Guidance on the responsible use of quantitative indicators in research assessment (DORA, 2024), it is essential to adopt a contextualized approach and enrich these indicators to ensure they effectively reflect the broader impact and quality of research.

Qualitative insights provide the necessary context that metrics alone cannot convey, highlighting nuances such as innovation, collaboration, and societal influence. The

narrative explanations of research activities and competencies, often culminating in the development of narrative curricula vitae (CVs). These narrative serve to provide a more comprehensive view of a researcher's accomplishments, contextualizing their contributions beyond numerical metrics. They highlight aspects such as interdisciplinary collaboration, innovation, and societal impact, which are difficult to quantify. However, researchers and evaluators identified potential drawbacks, including the significant burden on evaluators, who must process and critically analyze extensive qualitative descriptions. Furthermore, the subjective nature of narrative assessments introduces risks of bias proficiency in language, and cultural presentation skills. These elements can inadvertently favor individuals with stronger communication abilities or cultural capital, potentially leading to inequities in the evaluation process.

These qualitative insights, when combined with robust evidence, become a useful tool to enable fair and responsible research assessments. The development of the Researcher Profile takes this into account by embedding a qualitative perspective supported by quantitative information to provide a broader understanding of a researcher's contributions. In this context, this tool will contribute to a more responsible and comprehensive evaluation system by recognizing diverse research contributions beyond traditional metrics, fostering inclusivity, and ensuring a fairer assessment of researchers' work and societal impact.

An innovative strategy to support fair and responsible research assessment across disciplines will be the creation of different templates for the narrative CV and the profile itself, tailored to accommodate the diverse domains of researchers across scientific disciplines. Whether in engineering, social sciences, or other fields, our goal is to highlight researchers' contributions more effectively by providing discipline-specific formats that align with the nature of their work. This adaptability ensures that the profile remains relevant and equitable across various research areas, supporting a more inclusive and comprehensive assessment framework. By enabling flexibility in how achievements are presented and evaluated, this approach fosters a more nuanced and fair recognition of research excellence beyond traditional publication-based metrics.

Methodology: Designing the Framework for the Researcher Profile

The framework for the Researcher Profile aims to provide a customisable service that allows researchers to dynamically and seamlessly showcase their diverse contributions to research, knowledge and innovation. The design of the framework was based on a re-engineering process. Re-engineering in a scientific or engineering context refers to the process of redesigning or modifying existing systems, products, or technologies to improve performance, functionality, or to adapt to new requirements (Software Re-engineering: An Overview, 2018). Firstly, a landscape analysis of existing services and platforms (Google Scholar, Academia.edu, Web of Science, ResearcherID) was conducted, documenting their structure and the types of data they offer and showcase for researchers. Through this process, several pieces of information were gathered on the types of data presented and collected references to relevant indicators. Also key aspects of a researcher's career path, such as positions

held and extracurricular contributions and other activities about a researcher's career trajectory were collected.

To select novel indicators and categorize relevant activities for the Researcher Profile, the development of the service builds upon the ongoing work of Horizon Europe project OPUS (Open and Universal Science, 2022). More specifically, the OPUS project is working on a framework to assess researchers, including Open Science dimension to ensure that such practices are explicitly recognised and rewarded (O'Neill, 2023). From this framework, three main categories – "Research", "Education" and "Valorisation"– were identified in which the data collected, through our landscape analysis, were classified. These categories serve as a structured framework to organize and interpret the diverse information gathered from various sources.

Building on this foundation, ensuring a comprehensive evaluation that aligns with the UNESCO Recommendation on Open Science while leveraging insights from our landscape analysis Open Science as a fourth category was integrated. Open Science is essential for accelerating innovation, ensuring global access to knowledge, and fostering collaboration across disciplines, with many platforms from landscape analysis showcasing researcher's Open Science contributions. In the framework regarding the researcher's Open Science activities, the evaluation will include key indicators such as open access publishing and use of open-source software for research. However, the scope of Open Science will not be limited to these indicators; instead, it will be considered more holistically, encompassing key pillars as defined in the UNESCO Recommendation on Open Science (UNESCO, 2021).

To achieve the objectives outlined the required data on researchers' contributions will be sourced from ORCID and the OpenAIRE Graph. Specifically, information on researchers' education, qualifications and work experiences will be integrated from ORCID and automatically displayed within the developing service as recorded there. Similarly, data on research outputs, projects, and the researchers' broader network will be sourced from the OpenAIRE Graph, an extensive research database encompassing diverse research contributions. All collected information will be organized according to the initial prototypes developed through a design program. Moving forward, the aim is to integrate additional data sources to enhance the breadth and depth of the information collected, providing a more comprehensive view of researchers' contributions. Additionally, the plan is to implement a functionality that allows users to manually edit and update their data, ensuring flexibility and accuracy in maintaining profiles and related information. The aim of the developing service is to include all these diverse activities, showcasing the overall impact of the researcher's work. Finally, practical feedback on the components of the framework will be provided by the nine GraspOS Pilots (GraspOS – Pilots) who each represent a specific context in the research assessment system (National research funding and performing organisations, universities and university departments, disciplines).

Key characteristics of the Researcher Profile

The Researcher Profile service includes several key components, along with relevant indicators, to provide evidence of the researchers' contributions. Additionally, it outlines benchmarks and information that reflect the impact of the researchers' work across various dimensions. This holistic approach ensures that both qualitative and quantitative information are considered in research assessment.

A key element of this framework is the Narrative CV section, which will gather qualitative input on a researcher's skills and experiences. Based on the four-module model of the Royal Society's Résumé for Researchers (Resume for researchers), this approach supports a more contextual and qualitative assessment of their diverse contributions to research and society. These include contributions a) to the generation of knowledge, b) to the development of individuals, c) to the wider research community and d) to broader society. To further enhance the completeness of the profile, additional modules will present other types of experiences, such as extracurricular or voluntary work, thereby providing a more complete view of a researcher's curricular. This Narrative CV section will serve as the core feature of this profile, offering a comprehensive overview of achievements and contributions, providing context on the impact of their research, supported by evidence-like quantitative indicators. This section as referred above will differ and be tailored to accommodate the diverse domains of researchers across scientific disciplines. Complementing the Narrative CV, the interactive timeline will provide a dynamic, visual representation of a researcher's milestones, allowing users to explore the evolution of their research activities in a chronological order. By selecting on different elements, users can access more detailed information, making the exploration of data and narratives more engaging.

The Research Outputs section will further enrich the developing service by gathering a broad range of outputs including publications, preprints, datasets, and other research-related products, recognizing the need to showcase the variety of outputs produced in science. This section allows researchers to provide context through narrative boxes, enabling them to explain the rationale, activities, and outcomes behind their work.

Additionally, a dedicated section will recognize engagement with Open Science, highlighting researcher's activities who contribute to making scientific knowledge openly available, accessible, and reusable. The Researcher Profile aims to consider three pillars of Open Science, as reported in the UNESCO Recommendation on Open Science (UNESCO, 2021): open scientific knowledge, open science infrastructures, and open engagement of societal actors along with several Open Science indicators.

Valorisation will be another key feature, focusing on the broader impact of research. It refers to the process of enhancing the value of contributions often through refining, promoting, or developing it. This section will highlight how research contributes to practical applications, products, or social benefits, particularly in addressing real-world problems. By recognizing the societal and economic contributions of research, this section will complement the Researcher Profile focusing on the wider impact of research beyond academia.

Considerations for future developments and implementation

The main aim of the GraspOS project is to develop tools and services to support and facilitate the transition to Open Science-aware responsible research assessment practices. In light of the movement for reform and the growing emphasis on acknowledging a wider range of contributions to science and society, including Open Science practices, the framework under development promotes a balanced approach that combines qualitative information supported by quantitative indicators. However, as with any new service, the design and development of the Researcher Profile should carefully examine a variety of potential challenges.

There may be a risk that specific quantifiable Open Science practices or outputs substitute previous misused metrics, overlooking the need to monitor a comprehensive transformation of a new research culture. In addition, there is a need for assessing the values and impacts of science, focusing on the people who are conducting, engaging with, and/or benefiting from scientific work. Existing methods to assess the adoption of Open Science practices should therefore be strengthened (UNESCO, 2023), particularly to track the research culture change and value open and reproducible research processes.

The development of the Researcher Profile addresses several important considerations related to the flexibility of the developing profiles across diverse fields of study. One of the key challenges is ensuring that the framework can adapt to different contexts and needs across disciplines. Research contributions in fields such as the social sciences and applied sciences are fundamentally different in the way they are produced, disseminated, and evaluated.

A rigid, one-size-fits-all Researcher Profile structure would fail to capture the unique characteristics and impact of work in each domain, making it challenging for evaluators to assess the full breadth and depth of individual contributions. A unified approach would not only risk undervaluing important contributions, but it would also create unnecessary challenges for evaluators. The framework for the Researcher Profile needs to be flexible enough to be adapted to various local contexts and cater to research institutions' diverse values, needs and goals. Ultimately, the goal is to design a service that enables the creation of customizable, context-aware CVs, allowing researchers to highlight the achievements most relevant to their field.

Conclusion

The reform of research assessment is a critical step toward fostering a more inclusive, transparent, and Open Science-aware research culture. Traditional metrics, while providing useful benchmarks, have long been misapplied as proxies for quality, leading to systemic biases and misaligned incentives. In response to these challenges, the European policy agenda has prioritized the transition toward responsible research assessment, emphasizing the need for both qualitative and quantitative approaches that recognize diverse research contributions.

The Researcher Profile, developed as part of the Horizon Europe GraspOS project, offers a novel and pragmatic approach to addressing these challenges. By integrating qualitative insights with quantitative indicators, this innovative service ensures a

holistic and fair evaluation of researchers, acknowledging contributions beyond traditional publication-based metrics. The inclusion of a Narrative CV, interactive timeline, and domain-specific templates addresses the need for flexibility across disciplines, ensuring that researchers from all fields can effectively showcase their impact. Furthermore, the integration of Open Science principles aligns with international policy frameworks, reinforcing transparency, collaboration, and societal engagement in research assessment.

Looking ahead, the implementation of the Researcher Profile will require continuous refinement and collaboration with stakeholders to ensure its adaptability and effectiveness. It is imperative to prevent the replacement of old, flawed metrics with equally narrow Open Science indicators, instead fostering a cultural shift that values diverse research outputs and practices. As the research community advances toward a more responsible and comprehensive assessment system, the Researcher Profile serves as a pivotal tool in driving this transformation, ultimately strengthening the integrity, inclusivity, and impact of research within and beyond academia.

References

Declaration on Research Assessment (DORA). (2024). Guidance on the responsible use of quantitative indicators in research assessment. DORA. https://doi.org/10.5281/zenodo.10979644

Di Donato F., "What we talk about when we talk about research quality. A discussion on responsible research assessment and Open Science", Bollettino telematico di filosofia politica, March 2024. https://commentbfp.sp.unipi.it/quality-fdd/; https://doi.org/10.5281/zenodo.10890788

Edwards, M. A., & Roy, S. (2016). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science*, *34*(1), 51–61. https://doi.org/10.1089/ees.2016.0223

Elsevier Language Services (2020) *What is Journal Impact Factor?*, *Elsevier Author Services - Articles*. Elsevier Author Services – Blog. Available at: https://scientific-publishing.webshop.elsevier.com/research-process/what-journal-impact-factor/ (Accessed: March 26, 2025).

European Commission: Directorate-General for Research and Innovation. (2021). *Towards a reform of the research assessment system: scoping report*. Publications Office. https://data.europa.eu/doi/10.2777/707440.

European Commission: Directorate-General for Research and Innovation. (2022). *European Research Area policy agenda: overview of actions for the period 2022-2024*. Publications Office of the European Union. https://data.europa.eu/doi/10.2777/52110.

European Commission, Action Plan by the Commission to implement the ten commitments of the Agreement on Reforming Research Assessment (ARRA), 2024, https://research-and-innovation.ec.europa.eu/document/download/e69aff11-4494-4e5f-866c-694539a3ea26_en?filename=ec_rtd_commitments-reform-research-assessment.pdf

GraspOS open research assessment dataspace (no date) *About – GraspOS*. Available at: https://graspos.eu/. (Accessed: March 26, 2025).

GraspOS open research assessment dataspace (no date) *Pilots – GraspOS*. Available at: Pilots - GraspOS. (Accessed: March 26, 2025).

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*(7548), 429–431. https://doi.org/10.1038/520429a

International Network of Research Management Societies - Research Evaluation Group,. (2023, January 19). The SCOPE Framework. The University of Melbourne. Retrieved January 24, 2025, from https://figshare.unimelb.edu.au/articles/report/The_SCOPE_Framework/21919527/1

O'Neill, G. (2024). OPUS Deliverable 3.1: Indicators and Metrics to Test in the Pilots. Zenodo. https://doi.org/10.5281/zenodo.10497434

Open and Universal Science (2022) *Open and Universal Science (OPUS) Project - OPUS helps reform the assessment of #research in all research organisations towards a system that incentivise researchers to practice*. RAISE fosters startup growth and scale-up within and across Europe. Available at: https://opusproject.eu/. (Accessed: March 26, 2025)

Open science outlook 1: status and trends around the world. (2023). https://doi.org/10.54677/giic6829

Pontika, N., Klebel,T., Correia, A., Metzler, H., Knoth, P., Ross-Hellauer, T., Indicators of research quality, quantity, openness, and responsibility in institutional review, promotion, and tenure policies across seven countries. *Quantitative Science Studies*; 3 (4): 888–911, 2022. https://doi.org/10.1162/qss_a_00224

Résumé for researchers (no date) Royalsociety.org. Available at: https://royalsociety.org/news-resources/projects/research-culture/tools-for-support/resume-for-researchers/ (Accessed: March 26, 2025).

UNESCO Recommendation on Open Science. (2021). https://doi.org/10.54677/mnmh8546

Software Re-engineering: An Overview (2018). https://ieeexplore.ieee.org/abstract/document/8486173.

https://doi.org/10.51408/issi2025_166

Can scientific papers be unretracted?

Marek Kosmulski

m.kosmulski@pollub.pl
Lublin University of Technology, Nadbystrzycka 38, 20618 Lublin (Poland)

Introduction

Retraction of scientific papers makes it possible to "unpublish" a paper when a decision about its publication was premature. This is possible in electronic publishing, when the published items can be still edited, even long after their publication, as opposite to classical printed items, when the publisher cannot control the published items. A recent example shows that the decision about retraction can also be premature or at least debatable. What then? In principle the publisher can "unretract" a retracted article, that is, withdraw their decision about the retraction. However, retraction notes are indexed in scientific databases as WoS and Scopus, and archived, and the publisher cannot control these databases or archives. The other problem is that the decision about unretraction can also be premature (and so forth), and this may result in a loop of retractions and unretractions.

The citations of retracted papers and of retraction notes, and citations in retracted papers and in retraction notes are counted in scientific databases as the other citations, that is, it is not possible to automatically correct for them. The contribution of citations of retracted papers and of retraction notes, and of citations in retracted papers and in retraction notes to the total number of citations is negligible in large datasets, but at certain aggregation levels (e.g., in less successful journals and scientists), such a correction may have a substantial effect on the citation counts.

Case study

Machacek and Srcholec (2021) published a paper on predatory publishing in Scientometrics. Their paper was retracted (Machacek and Srcholec, 2022) by the Editor-in-Chief. The retracted article received 49 citations (April 2025), which is well above the average in scientometrics and in Scientometrics.

A group of outstanding bibliometricians (Abramo et al., 2023) criticized the decision about the retraction, and received an answer from the Editor-in-Chief (Zhang, 2023). In the meantime Machacek and Srcholec (2022a) re-published their retracted article in another journal, and the new version received further 20 citations. Some authors citing the re-published version might not be aware of the original (retracted) version. In contrast most authors who recently cited the original version were aware that they cited an retracted article, because the availability of printed journals is limited, and most scientist use the articles loaded from Internet, where retracted articles are clearly marked as such.

Most citations of Machacek and Srcholec (2021) and of Machacek and Srcholec (2022a) refer to the substance of their article(s?) and only a few of them refer to the very fact that the article was retracted.

On top of discussions in journal articles, the retraction was discussed by Retraction Watch, and in two items, which look like short papers in Scientometrics (published on the Web page of Scientometrics), but they do not have volume or page numbers.

Discussion

Predatory journals discussed by Machacek and Srcholec in their retracted paper are a sensitive topic, and obviously the editors and publishers of journals deemed predatory will protest against such a classification of their journals. There is neither commonly accepted definition of predatory journals not a sharp border between predatory and non- predatory journals. The history of Beall's list is the most well-known example of such an attitude of editors and publishers.

Due to the touchiness of the topic, the editorial decisions with respect to papers on predatory journals should be made with a special care including the anticipation that someone will feel offended by the publication. The topic of

predatory journals is not unique in this respect. There are numerous equally sensitive topics in health care, religion science, ecology, etc.

Touchy topics cannot be completely avoided in science. Especially the predatory publishing became an essential part of the scientific landscape.

Two papers by Machacek and Srcholec created a dangerous precedent: two scientific papers of the same authors, under the same title and with basically the same substance. This situation could have been avoided when the editor had a chance of having withdrawn their decision about subtraction, for example after reaction of other scientists to subtraction, as described in the above case study.

Study in progress

This is not clear if the above story of the paper by Machacek and Srcholec, that is, re-publication of the once retracted article in another journal, is unique, or more examples like this can be found. The study is in progress.

References

Abramo, G., Aguillo, I.F., Aksnes, D.W. et al. (2023) Retraction of Predatory publishing in Scopus: evidence on cross-country differences lacks justification. *Scientometrics* 128, 1459–1461. https://doi.org/10.1007/s11192-022-04565-6.

Macháček, V., & Srholec, M. (2021). Predatory publishing in Scopus: Evidence on cross-country differences. *Scientometrics*, 126, 1897–1921. https://doi.org/10.1007/s11192-020-03852-4.

Macháček, V., & Srholec, M. (2022). Retraction note to: Predatory publishing in Scopus: Evidence on cross-country differences. *Scientometrics*, 127, 1667. https://doi.org/10.1007/s11192-021-04149-w.

Macháček, V., & Srholec, M. (2022a). Predatory publishing in Scopus: Evidence on cross-country differences. *Quantitative Science Studies* 3, 859–887. https://doi.org/10.1162/qss_a_00213.

Zhang, L. (2023) Editorial response letter to Abramo et al. Scientometrics, 2022. *Scientometrics* **128**, 1463–1464. https://doi.org/10.1007/s11192-022-04608-y.

https://doi.org/10.51408/issi2025_167

Ethical and responsible model for the National Science, Technology and Innovation System in Colombia

María Alejandra Tejada-Gómez[1], Mabel Ayure-Urrego[2]

[1] *maria.tejada@javeriana.edu.co*
Asesora Vicerrectoría de Investigación, Pontificia Universidad Javeriana, Bogotá (Colombia)

[2] *mabelayure@gmail.com*
Secretaría de Educación del Distrito de Bogotá (Colombia)

Introduction

Colombia's National Science, Technology, and Innovation System (NST&IS) faces important challenges in advancing towards a responsible and ethical measurement of research results. In 2023, a comprehensive model was designed for the modernization of System, based on the studies developed by Tejada-Gómez (2022) and Ayure-Urrego (2021). This model focuses on fundamental values for the scientific community, such as trust, integrity, and responsibility (Figure 1).

Dimensions of the Modernization Model

The Modernization Model proposed to address the following dimensions:

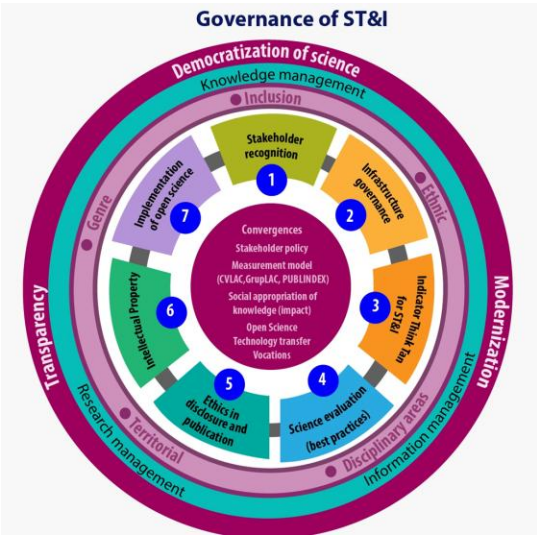


Figure 1. Modernization Model for the NST&IS of Colombia.
Developed by Tejada-Gómez and Ayure-Urrego (2023).

Recognition of the types of actors involved in R&D activities.

The model for Modernizing the National Science, Technology and Innovation System integrates the management of research, information, and knowledge from an inclusive approach, with a gender, ethnic, and territorial perspective.

Based on current science policy instruments and UNESCO and OECD recommendations, to progressively involve citizens and vulnerable or minority communities in the development of science.

Governance of the infrastructure for scientific information.

Colombia has an information platform for organizations, researchers, and research results that requires integrating current scientific information management resources, such as persistent identifiers, and moving towards an interoperable model.

R&D Indicators Think Tank

The characteristics of the country require a data and indicators model with territorial and inclusive details, likewise, the official sources that measure, calculate, and report data/indicators require collaborative and public governance instruments.

Research evaluation

The current models for evaluation require flexibility in the types of research results, knowledge products, and ways of acting in the development of science. The evaluation path must also promote changes in the regulatory instruments that define researchers' careers,

the forms of participation of R&D organizations, and open science.

Ethics of research dissemination and publication

One of the most sensitive aspects is the flexibility of the model for the recognition and measurement of scientific publications. Although Colombia has official policies for ethics and scientific integrity, it is necessary to advance in the transformation of the current mechanisms for the valuation of scientific publications. Likewise, it is necessary to advance globally in the aspects of public communication of science, which gives rise to the participation of more actors involved in R&D activities.

Intellectual Property Rights

Recognition of intellectual property rights and progress in the democratization of science are fundamental for scientific careers and trajectories, as well as the transfer of knowledge to promote innovation.

Progress and collaborative work

As a result of the collaborative work between national and international technical experts, progress was made in the following areas:

- Scientific information infrastructures
- Measurement model for national scientific publications
- Measurement model for research
- National system of indicators for the measurement of research
- Ethics for measurement and evaluation in the Publindex Model.

The progress in the dimensions of the modernization model for the Colombian SNCTeI showed the need to make efforts on the part of the organizations that carry out R&D activities, government decision makers, science policy makers, funders and other allies, to carry out integral processes of research and measurement of research results with multidimensional and inclusive views, based on ethics and responsibility about the characteristics and national and global context. In the coming years, decision makers, public policy makers, and main funders of scientific and technological research and innovation processes should ensure the transition to:

- Ethical and multidimensional models for the collection, management, and use of national research data and indicators.
- Sustainable and public infrastructures to manage data and indicators of the Colombian SNCTeI.
- Inclusive evaluation systems align with COARA.

References

Ayure-Urrego, M. (2021). *Prácticas de Comunicación Pública de Ciencia y Tecnología en museos de ciencia. Parque Explora (Medellín) y Cosmo Caixa (Barcelona)*. Universidad de Barcelona.

COARA (2022). Agreement on reforming research assessment. Coalition for Advancing Research Assessment

Curry, Stephen; Gadd, Elizabeth; Wilsdon, James (2022). Harnessing the Metric Tide: indicators, infrastructures & priorities for UK responsible research assessment. Research on the Research Institute. Report.

https://doi.org/10.6084/m9.figshare.21701624.v2

FOLEC (2021). Toward the transformation of evaluation systems in Latin America and the Caribbean. Tools to promote new evaluation policies.

Himanen L., Conte E., Gauffriau M. et al. The SCOPE Framework – Implementing the ideals of responsible research assessment [version 1; peer review: awaiting peer review] F1000Research 2023, 12:1241

OECD (2023). Developing an Implementation Toolkit for the Recommendation of the OECD Council concerning Access to Research Data from Public Funding. Directorate for science, technology, and innovation committee for scientific and technological policy. DSTI/STP (2023)15, 3 March 2023

OECD (2023), "Science, technology and innovation policy in times of global crises", in OECD Science, Technology and Innovation Outlook 2023: Enabling Transitions in Times of Disruption, OECD Publishing,

Paris, https://doi.org/10.1787/d54e7884-en.

OECD (2024), "OECD Agenda for Transformative Science, Technology and Innovation Policies", *OECD Science, Technology and Industry Policy Papers*, No. 164, OECD Publishing, Paris, https://doi.org/10.1787/ba2aaf7b-en.

Chalela Naffah Salim; Corral Strassman Maria Mercedes; Lucio-Arias Diana; Pallares Delgado César Orlando; Tejada-Gómez María Alejandra (2023). Definición responsable de métricas para la evaluación de la investigación en Colombia. Asociación Colombiana de Universidades ASCUN, Observatorio Colombiano de Ciencia y Tecnología.

Tejada-Gómez, M. A. (2022). University Research Governance and the Colombian Scientific Journal Index "Publindex:" Understanding the Tensions. Twente University https://doi.org/10.3990/1.9789036553698

Science Europe (2020). Position Statement and Recommendations on Research Assessment Processes.

Wilsdom, J., et al. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. DOI: 10.13140/RG.2.1.4929.1363

Special Track: Open Research Information (ORI)

https://doi.org/10.51408/issi2025_150

Usage, Advancements, and Limitations of Open Research Information Data Sources

An Introduction to the ISSI 2025 Special Track "Open Research Information" (ORI)

Robin Haunschild[1], Lin Zhang[2], Erjia Yan[3]

[1]*r.haunschild@fkf.mpg.de,* ORCID: 0000-0001-7025-7256
IVS-CPT, Max Planck Institute for Solid State Research, Germany
2linzhang1117@whu.edu.cn, ORCID: 0000-0003-0526-9677
Wuhan University, Hubei, China
[3]*erjia.yan@drexel.edu;* ORCID: 0000-0002-0365-9340
Drexel University, PA, USA

Open research information is information on scientific research that is freely available to the public to access, use, and reuse. This includes but is not limited to bibliographic data and metadata regarding research publications, software, tools, and information about research processes like funding and project details. Open research information has begun to transform the way research is conducted and its results are published. By being openly accessible, open research information aims to promote transparency, reproducibility, collaboration, and innovation in the research community. Open research information not only fosters a culture of reproducibility but also a culture of accountability and public engagement in science. It also fuels innovative research, enabling researchers, institutions, policymakers, educators, and the public to freely access, use, and build upon scientific knowledge and thereby advancing research and its societal impact.

Open research information, its technical infrastructures and social architectures that support management, delivery, and preservation of research information, constitute a bedrock of open science. It acts as shared resources, slicing across disciplinary and geographic boundaries, benefiting stakeholders and constituents in the scientific ecosystem. Open research information forms a critical pillar of the open science movement and transformation, especially in light of the Barcelona Declaration on Open Research Information proclaimed in April 2024. However, it also presents significant challenges in terms of management, curation, design, maintenance, governance, ethics, sustainability, and impact measurement. Recognizing both the vast potential and the complexities, as well as the challenges inherent in open research information, we are happy to organize a special session on open research information at ISSI 2025. The role of bibliometric indicators in research evaluation has undergone substantial evolution over the past decades, becoming integral to institutional assessments, funding decisions, and science policy at large. While their widespread adoption has enabled new forms of analysis and benchmarking, it has also sparked ongoing debates around their transparency, ethical integrity, and contextual relevance. A central concern is the over-reliance on narrow performance metrics, such as publication counts or citation-based rankings, often applied

uniformly and without sufficient consideration of disciplinary norms, research diversity, or the broader societal value of scientific work.

Several aspects need to be considered to levy the full potential of open research information. The following aspects are discussed in this special track:

1. Comparison of the coverage of open research information data sources in comparison to local repositories.
2. Combination of different open research information data sources.
3. Application of data-mining techniques to the data inside open research information data sources.

At the outset of the special track "Open Research Information" hosted at ISSI 2025, we present an overview summarizing recent advancements in open research information data sources (Cao, Zhang, Huang, and Haunschild, 2025) as an introduction into the special track. Following this, we invite the presenters of the following selected papers included in the track, each of which addresses key aspects of the aforementioned topic:

- "What Are We Missing? A Systematic Approach to Overlap Analyses of Local and Global Repositories" (Willemin, Bernard, Dederke, Hemila, and Koch, 2025).
- "How well does OpenAlex cover the Flemish Social Sciences and Humanities?" (Vandewalle and Arhiliuc, 2025).
- "Multi-Disciplinal, Large Scale Mentorship Dataset and Demographics" (Miura, Watanabe, Sakammoto, and Hashizume, 2025).
- "Annotation and Identification of Scientific Data Sharing Information from Data Availability Section" (Xu, Li, An, Wang, Li, and Zhang, 2025).

By assembling selected contribution to three particular aspects relevant to Open Research Information Data Sources and encouraging a lively discussion, this special track contributes to pointing to limitations and providing solutions to problems as well as leveraging the advantages of Open Research Information Data Sources.

List of contributions to the Special Track

Cao Z., Zhang L., Huang Y., and Haunschild R. (2025), "How does the academia refer to open research information data sources? A review study based on OpenAlex and Microsoft Academic series", Scientometrics (in press).

Miura C., Watanabe Y., Sakammoto T., Hashizume H. (2025), "Multi-Disciplinal, Large Scale Mentorship Dataset and Demographics", in this Special Track.

Vandewalle E. and Arhiliuc C. (2025), "How well does OpenAlex cover the Flemish Social Sciences and Humanities?", in this Special Track.

Willemin S., Bernard G., Dederke J., Hemila M., and Koch M. (2025), "What Are We Missing? A Systematic Approach to Overlap Analyses of Local and Global Repositories", in this Special Track.

Xu S., Li J., An X., Wang S., Li J., Zhang Y. (2025), "Annotation and Identification of Scientific Data Sharing Information from Data Availability Section", in this Special Track.

https://doi.org/10.51408/issi2025_083

Annotation and Identification of Scientific Data Sharing Information from *Data Availability* Section

Shuo Xu[1], Jiahao Li[2], Xin An[3], Shengnan Wang[4], Jianhua Liu[5], Yuefu Zhang[6]

[1] *xushuo@bjut.edu.cn*, [2] *lijh0707@emails.bjut.edu.cn*, [6] *yaogeng_z@163.com*
School of Economics and Management, Beijing University of Technology, Beijing (China)

[3] *anxin@bjfu.edu.cn (Corresponding author)*, [4] *18706438326@163.com*
School of Economics and Management, Beijing Forestry University, Beijing (China)

[5] *liujh@wanfangdata.com.cn*
Beijing Wanfang Data Co., LTD, Beijing (China)

Abstract

With the advancement of the open science movement, an increasing number of institutions and journals now require authors to explicitly state data availability in their publications, thus promoting the open sharing and accessibility of scientific data. The aim of this study is to extract scientific data sharing information from data availability statements in scientific papers. In more detail, this study annotates 8,508 data availability statements in research papers from the PLOS corpus over a period of nearly 16 months. In the end, a total of 35,010 entities and 8,524 relations covering 8 types of entities and 2 types of semantic relationships are ultimately annotated. Based on the annotated data, the model on the basis of Universal Information Extraction (UIE) is fine-tuned to automatically identify entity and relation mentions from data availability statements of the remaining scholarly articles. Experimental results show that our model is capable of extracting scientific data sharing information.

Introduction

With the development of open science movement, the open sharing of scientific data has progressively become a significant trend (Xu et al., 2021; Lu et al., 2024). Numerous countries and funding organizations worldwide have actively implemented policies to promote the public availability and standardized management of data (Jiao, Qiu, Ma, & Yang, 2024). In the context of increasing emphasis on the openness and transparency of research data, the emergence of data sharing information within scientific data statements has further laid the groundwork for the standardization and institutionalization of data sharing practices (Yang, Zhang, & Huang, 2023). By data sharing information within scientific data statements, we mean the declarations within scientific publications that outline how scientific data is stored, shared, and accessed.

To enhance the digital ecosystem of scientific data in the process of open sharing, Wilkinson et al. (2016) systematically introduced and defined the FAIR (i.e., Findable, Accessible, Interoperable, and Reusable) principles, which provide an internationally recognized framework for the management and sharing of scientific data. Correspondingly, all journals published by PLOS [1] and Springer Nature [2]

[1] https://journals.plos.org/plosone/s/data-availability
[2] https://www.springernature.com/gp/authors/research-data-policy

issued new open data policies. The submission guidelines explicitly state that all scientific data supporting conclusions must be stored in public data repositories that comply with FAIR principles and provide corresponding DOIs or access numbers. Additionally, the data availability statement must clearly outline any access restrictions or special conditions, such as limitations due to legal or ethical constraints or the requirement of an application for access. These statements are usually located in *Data Availability* section. This enables the accessibility and evaluation of scientific data sharing information at large scale.

Federer et al. (2018) collected data availability statements from the articles published in PLOS ONE journal between March 2014 and May 2016, and found that only approximately 20% of the statements indicated the data were stored in a repository. After then, the long-term availability of URLs and DOIs mentioned in the data availability statements of PLOS ONE articles were further examined. Federer (2022) observed that approximately 80% of the resources could be successfully retrieved, whereas the retrieval rate relying on author contact to locate data was substantially lower, ranging from 10% to 40%. Subsequently, Jiao et al. (2024) took into consideration the articles published in PLOS ONE journal from 2014 to 2020, and employed the rules on the basis of regular expressions to extract data sharing mechanisms and repositories from the data availability statements. Jiao et al. (2024) argued that although data continued to be primarily shared through the main article or its supplementary materials, the use of data repositories exhibited a steady growth trend.

It is not difficult to see that previous studies are just limited to the articles published in PLOS ONE journal. In addition, since sharing information often appears in the form of diverse and irregular expressions, this results in unsatisfactory performance in sharing information extraction with rule-based approaches. Hence, this study considers all the articles published in the journals by PLOS publisher, and annotates a large-scale and high-quality dataset for data sharing information, encompassing eight types of entities and two types of semantic relationships. What's more, an automated identification model is constructed with the help of Universal Information Extraction (UIE) (Lu et al., 2022).

Data Annotation

Data sources

Since 1 March 2014, PLOS has implemented a data availability policy, requiring all submitted manuscripts to provide a detailed description of data sharing compliance within the data availability statement (Bloom, Ganley, & Winker, 2014). Therefore, the PLOS corpus [3] is selected as the data source in this study. This corpus was downloaded on August 21, 2023, comprising a total of 338,810 papers (excluding *correction* and *expression of concern* articles), in which 189,369 papers are attached with a section of data availability statements. On preliminary

[3] https://plos.org/text-and-data-mining/

analysis, we observe that many data availability statements are very simple, such as "All relevant data are within the paper and its Supporting Information files.", and "All relevant data are within the paper." As for these cases, several rules based on regular expressions are manually curated to match 107,747 scientific publications. In this way, 81,622 articles remain, from which 8,508 ones are randomly drawn for annotating entities and semantic relationships.

Definition of Entities and Relations

This study defines 8 types of entities: DATASET_NAME, ACCESS_NUMBER, REPOSITORY_FROM, REPOSITORY_TO, HREF_FROM, HREF_TO, TELEPHONE, and EMAIL, along with 2 types of relationships: SPAN and SAME_AS.

An example of data availability statements with annotated entity and relation mentions is illustrated in Figure 1. From Figure 1, it is easy to understand the DATASET_NAME, ACCESS_NUMBER, TELEPHONE, EMAIL, and SPAN. As for REPOSITORY_FROM, REPOSITORY_TO, HREF_FROM, and HREF_TO, the suffix "FROM/TO" can distinguish between data source repositories/hyper-references and data storage repositories/hyper-references. The semantic relation SAME_AS is mainly used to establish clear and standardized connections between different repository or URL mentions. Note that the SAME_AS holds between the entities with the following types: REPOSITORY_FROM vs. REPOSITORY_FROM, HREF_FROM vs. REPOSITORY_FROM, REPOSITORY_TO vs. REPOSITORY_TO, and HREF_TO vs. REPOSITORY_TO.

Figure 1. An example of data availability statements with annotated entity and relation mentions (DOI = "10.1371/JOURNAL.PONE.0171481").

Data labeling

The data annotation process is empowered by the web-based annotation tool BRAT (Wang et al., 2023). Our team consists of 3 members with one team leader (the first author of this work), and spends approximately 16 months. To ensure consistency, the team strictly manages the online+offine annotation process. First, before annotating, all annotators are trained. Second, the team leader regularly conducts sample audits of the annotation results and provides corrections and guidance for typical errors. Finally, the workload of each annotator is adjusted periodically based on their annotation results. The annotators with lower accuracy experience a corresponding reduction in their workload.

Throughout the annotation process, three to four rounds of refinement are involved. After each round is completed, all annotated mentions are reviewed by our team leader, the resulting feedbacks are incorporated to optimize and adjust the annotation guidelines. Taking REPOSITORY as an example, the annotation guidelines underwent the following changes: In the first iteration, we focus on annotating repositories in the papers that explicitly mention data storage locations, with popular repositories such as Figshare, the NCBI database, and the Genbank database. In the second iteration, the rules for ethics committees are added. If a paper references an ethics committee providing dataset access, such as "Requests for access to the data should be made to the Medical Ethics Committee of the Second Affiliated Hospital of Nantong University," this entity mention should be annotated. In the third iteration, the annotation guidelines for organizations are introduced. In this case, the organizations related to data requests are considered. For instance, in the statement "The data are not publicly available owing to privacy or ethical restrictions, as they contain sensitive information. The data are held by the Anhui Provincial Tuberculosis Institute. Requests to access the data can be sent to Xiao-Hong Kan, Chief of Science and Education at the Anhui Provincial Tuberculosis Institute." where it is explicitly stated that data requests should be directed to Anhui Provincial Tuberculosis Institute, this entity should be annotated. Finally, in the fourth iteration, the annotation rules are established for supplemental files. In the end, a total of 35,010 entity mentions and 8,524 relation mentions are annotated. The distribution is illustrated in Figure 2.

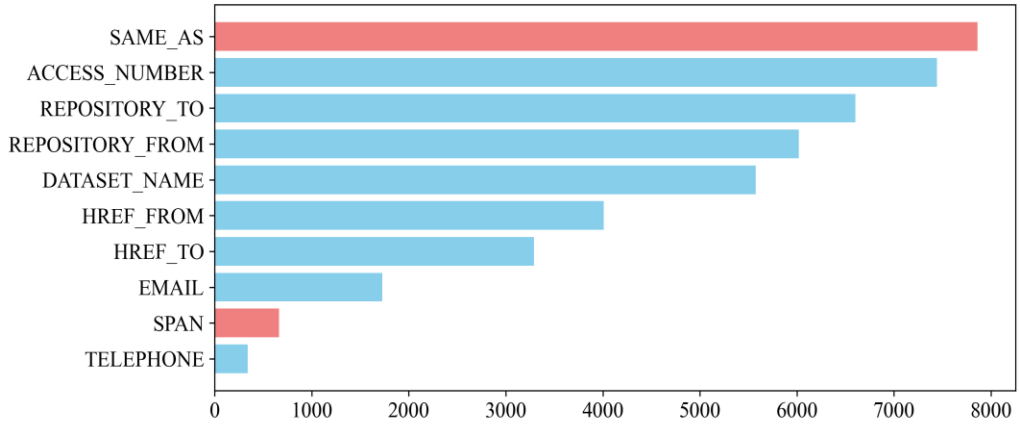


Figure 2. Distribution of number of entity and relation mentions in the annotated dataset.

As observed in Figure 2, several key characteristics can be observed as follows. (1) ACCESS_NUMBER (7,442) has the highest annotation count. This indicates that data access numbers are most frequently referenced in the data availability statements of research papers, highlighting their central role in data sharing. (2) The relatively high annotation frequency of REPOSITORY_FROM and REPOSITORY_TO suggests that the formal storage and traceability of scientific data are of significant concern in research. Notably, REPOSITORY (6,604) is annotated far more frequently than HREF (3,289), reflecting a tendency among researchers to directly reference data storage platforms or databases rather than individual web links. (3) The relatively low annotation frequencies of EMAIL (1,727) and TELEPHONE (340) suggest that instances of restricted data access still exist, albeit to a limited extent.

Entity and Relation Mentions Recognization Framework

The UIE framework

UIE (Universal Information Extraction) (Lu et al., 2022) represents a comprehensive framework for information extraction. Based on this framework, the PaddleNLP has developed and open-sourced the inaugural UIE model, with the ERNIE 3.0 as knowledge-enhanced pre-trained architecture. This model exhibits significant advantages in cross-domain adaptability, few-shot fine-tuning and efficient task transfer. More notably, UIE provides strong support for customizable model fine-tuning, allowing one to further refine the model using domain-specific data to optimize its performance in specialized fields or tasks.

In this study, the extraction of scientific data sharing information primarily involves two tasks: entity recognition and relation extraction. Traditional approaches (Chen et al., 2020) typically necessitate the independent training of two separate models, which significantly increases training complexity and may lead to a loss in predictive accuracy. In contrast, the UIE, by sharing network parameters, enables both tasks to be handled simultaneously within a unified framework,

reducing computational redundancy and resource waste. Moreover, UIE offers enhanced flexibility and scalability, making it more suitable for addressing complex application scenarios like our case.

UIE Model Fine-tuning

When training and inference are based on the BERT model, the maximum length of each input text is typically limited to 512 tokens (Xu et al., 2024), a constraint inherent to its architectural design. Since the UIE utilizes BERT as the underlying pre-trained model, it is similarly constrained by input length during the fine-tuning process (defaulting to 512 tokens). While the length can be extended, doing so significantly increases the consumption of computational resources.

This study further analyzes the textual characteristics of PLOS corpus. It is found that most samples adhere to the 512-token limit, although a subset of texts exceeds this length. To enhance the capacity to handle long texts, we select 786 tokens as the maximum input length for fine-tuning, taking into account both the input limitations of the pre-trained model and computational costs. This length accommodates the majority of samples, minimizes the loss of information due to excessive truncation, and improves the model's understanding of long texts. In more detail, 7,441 samples have a text length not exceeding 786 tokens in our annotated dataset.

To ensure consistency, we exclude the samples with text length more than this limitation in the training phrase. During model training, the dataset is further split into training, validation, and test sets in a 7:2:1 ratio, with 5,209 samples used for training, 1,488 for validation, and 744 for testing. Samples with a text length exceeding 786 tokens total 1,067. To handle the samples with text length more than this limitation, we employ a sliding window approach for segmentation and evaluation. Specifically, a fixed-size window (Xu et al., 2024) is applied to the original text, with a certain overlap maintained during each segmentation. This method ensures that more coherent semantic information is captured when processing long texts.

During the fine-tuning process, the UIE model demonstrates strong entity recognition capabilities on both the validation and test sets. As shown in Table 1, our model performs well on most entity types in terms of Precision, Recall, and F1-score. EMAIL and TELEPHONE nearly achieved perfect recognition performance, while entity types such as ACCESS_NUMBER, REPOSITORY_TO, HREF_FROM, and HREF_TO also maintained evaluation scores above 0.95, reflecting excellent recognition performance. However, the UIE model demonstrated relatively weaker performance on DATASET_NAME and REPOSITORY_FROM, particularly in terms of Recall. In our opinion, this issue is partly related to the nature of the entities in the data availability statements themselves. For instance, it is usually difficult to determine the connotation and denotation of a dataset name. This enables many annotated entity mentions with the DATASET_NAME category not to always point to a publicly available dataset name, such as "raw metagenomic sequencing data".

A similar issue is observed in long texts. As shown in Table 1, although the overall prediction accuracy remains at a commendable level, certain entity types, such as DATASET_NAME and ACCESS_NUMBER, exhibit a noticeable decline in terms of Precision and F1-score. This indicates that while the UIE demonstrates the capacity to some extent for handling long texts, its generalization ability may be limited in cases involving complex information.

In the relation extraction task, SPAN benefits from its clear structural characteristics, maintaining strong recognition performance in both short and long texts. In contrast, SAME_AS involves more complex structure and a wider range of entity types, which increases the difficulty of relation extraction. Specifically, in long texts, where more intricate contextual information and potential ambiguities arise, SAME_AS faces greater challenges.

Table 1. Evaluation Performance of UIE Model on the Validation Set / Test Set / Long Texts.

	Precision	Recall	F1-score
DATASET_NAME	0.8133 / 0.8449 / 0.5947	0.6657 / 0.6765 / 0.6708	0.7321 / 0.7321 / 0.6304
ACCESS_NUMBER	0.9852 / 0.9879 / 0.6698	0.9926 / 0.9712 / 0.9721	0.9889 / 0.9795 / 0.7931
REPOSITORY_FROM	0.8725 / 0.8720 / 0.7583	0.7802 / 0.7967 / 0.7555	0.8238 / 0.8326 / 0.7569
REPOSITORY_TO	0.9602 / 0.9526 / 0.8812	0.9468 / 0.9393 / 0.8892	0.9534 / 0.9459 / 0.8852
HREF_FROM	0.9939 / 0.9867 / 0.7842	0.9290 / 0.8810 / 0.9462	0.9604 / 0.9308 / 0.8576
HREF_TO	0.9871 / 0.9955 / 0.7610	0.9147 / 0.8975 / 0.8118	0.9495 / 0.9440 / 0.7856
TELEPHONE	1.0000 / 1.0000 / 0.8818	0.9756 / 1.0000 / 0.9418	0.9877 / 1.0000 / 0.9108
EMAIL	1.0000 / 1.0000 / 0.8682	1.0000 / 1.0000 / 1.0000	1.0000 / 1.0000 / 0.9295
SPAN	0.9806 / 0.9831 / 1.0000	0.9712 / 0.9831 / 0.9730	0.9759 / 0.9831 / 0.9863
SAME_AS	0.9250 / 0.8889 / 0.8443	0.8216 / 0.8085 / 0.7030	0.8703 / 0.8468 / 0.7672

Model Prediction and Analysis

During the model prediction phase, we primarily focus on the data availability sections of the remaining 73,114 papers. As shown in Figure 3，the identification results exhibit a pronounced long-tail distribution. Among the entities, REPOSITORY_TO (139,774) exhibits the highest frequency, emphasizing the central role of repository storage in data sharing. REPOSITORY_FROM (48,973) follows, slightly surpassing ACCESS_NUMBER (47,532). Entities with moderate

frequencies include DATASET_NAME (37,129), HREF_TO (31,129), and HREF_FROM (26,944). Among the relation types, SAME_AS (51,223) dominates.

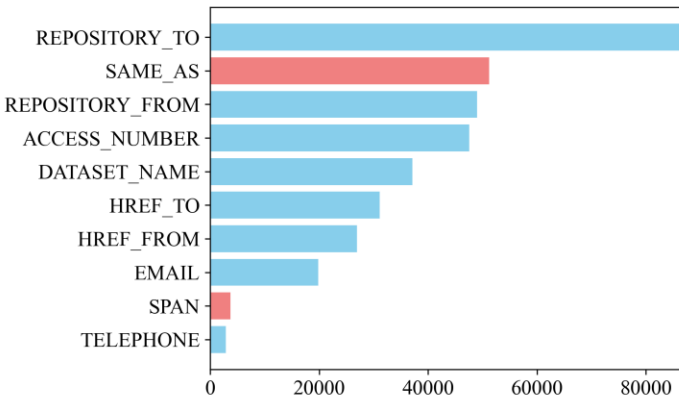
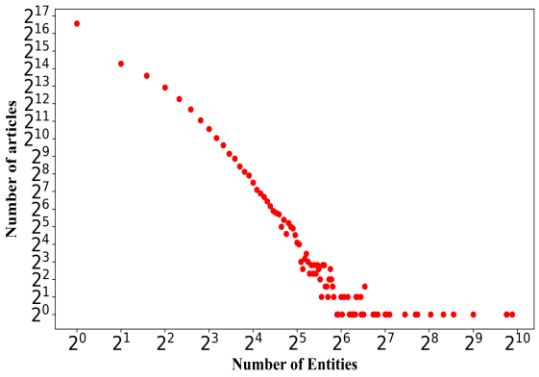
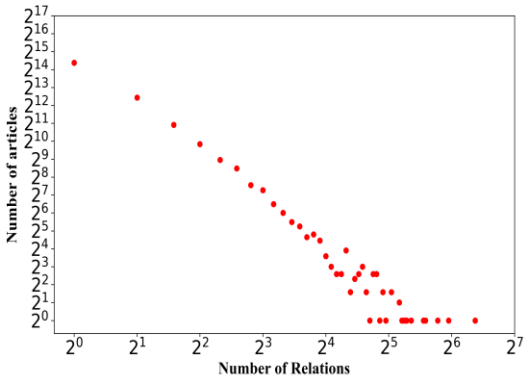


Figure 3. Distribution of entities and relations in the predicted dataset.

The log-log curves in Figures 4(a) and 4(b) illustrate the relations between the number of articles and the number of entity mentions, and the number of articles and the number of relation mentions, respectively. As the number of entity/relation mentions increases, the number of articles follows a typical power-law trend. To say it in another way, most articles contain fewer entities or relations, while articles containing a large number of entities or relations are relatively rare.





Figur4. Log-log curve of the number of articles and the number of entities (a), and the number of articles and the number of relations (b).
(a) Log-log curve between the number of articles and the number of entities
(b) Log-log curve between the number of articles and the number of relations

Conclusions and Limitations

In the context of the growing openness and transparency of scientific data, data availability statements, as one of the primary means of data sharing, have been

widely implemented and received significant attention across various academic journals. Previous studies primarily focused on the articles in PLOS ONE journal, rule-based approaches were usually resorted for extracting shared information, resulting in an unsatisfactory performance.

Therefore, this study randomly selects 8,508 articles published in the journals by PLOS publisher for the annotation of entities and semantic relationships. Through rigorous multiple rounds of manual annotation and quality review, this study ultimately constructs a high-quality corpus containing 8 types of entities and 2 types of semantic relationships, with a total of 35,010 entity mentions and 8,524 relation ones. Building on this, the study fine-tunes a model based on the UIE information extraction framework to achieve automated identification of entities and relations.

Though, there is still some room to improve our study as follows. The UIE framework under-performs when handling low-frequency entity types and relationships with ambiguous boundaries. Moreover, the performance in processing long texts needs to be further improved.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant numbers 72474016, 72004012 and 72074014).

References

Bloom, T., Ganley, E., & Winker, M. (2014). Data access for the open access literature: PLOS's data policy. *PLoS Medicine*, 11(2), e1001607.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., & Yang, G. (2020). A deep learning based method for extracting semantic information from patent documents. *Scientometrics*, 125(1), 289-312.

Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y. L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of data availability statements. *PloS ONE*, 13(5), e0194768.

Federer, L. M. (2022). Long-term availability of data associated with articles in PLOS ONE. *PloS ONE*, 17(8), e0272845.

Jiao, C., Li, K., & Fang, Z. (2024). Data sharing practices across knowledge domains: A dynamic examination of data availability statements in PLOS ONE publications. *Journal of Information Science*, 50(3), 673-689.

Jiao, H., Qiu, Y., Ma, X., & Yang, B. (2024). Dissemination effect of data papers on scientific datasets. *Journal of the Association for Information Science and Technology*, 75(2), 115-131.

Lu, L., Zhong, Y., Luo, S., Liu, S., Xiao, Z., Ding, J., ... & Xu, J. (2024). Dilemmas and prospects of artificial intelligence technology in the data management of medical informatization in China: A new perspective on SPRAY-type AI applications. *Health Informatics Journal*, 30(2), 14604582241262961.

Lu, Y., Liu, Q., Dai, D., Xiao, X., Lin, H., Han, X., ... & Wu, H. (2022). Unified structure generation for universal information extraction. *arXiv preprint arXiv*:2203.12277.

Wang, Z., Xu, S., Wang, Y., Chai, X., & Chen, L. (2023). Bureau for rapid annotation tool: Collaboration can do more among variance annotations. *Aslib Journal of Information Management*, 75(3): 523-534.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1-9.

Xu, S., Zhang, Y., Chen, L., & An, X. (2024). Is metadata of articles about COVID-19 enough for multi-label topic classification task? *Database*, 2024, baae106

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., ... & Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation*, 2(4), 100179.

Yang, N., Zhang, Z., & Huang, F. (2023). A study of BERT-based methods for formal citation identification of scientific data. *Scientometrics*, 128(11), 5865-5881.

https://doi.org/10.51408/issi2025_074

How well does OpenAlex cover the Flemish Social Sciences and Humanities?

Eline Vandewalle[1], Cristina Arhiliuc[2]

[1]*eline.vandewalle@uantwerpen.be*, [2]*cristina.arhiliuc@uantwerpen.be*
Centre for Research and Development Monitoring (ECOOM), University of Antwerp,
Middelheimlaan 1, 2020 Antwerp (Belgium)

Abstract

Since the launch of OpenAlex as a fully open and non-proprietary alternative to bibliographic indexing services, interest has risen in the extent to which OpenAlex covers the research landscape and in what areas it could increase coverage compared to the proprietary alternatives, particularly of the social sciences and humanities (SSH) and for publications in languages other than English. In this study, we have used the VABB-SHW database as a benchmark to compare OpenAlex with. VABB-SHW is a local comprehensive bibliographic database for the SSH. It includes many Dutch-language publications, and non-article publication types. We find that OpenAlex covers 50.46% of publications from the local bibliographic database (both peer-reviewed and non-peer-reviewed publications), with higher percentages for publications that are also indexed in the Web of Science (94.51%). Coverage is lower for non-English language publications and publication types other than articles. Additionally, we explore the metadata coverage in OpenAlex and find that 86 percent of the publications found in OpenAlex have reference data available and 91 percent of them have affiliation information. We also report on the strategy for matching records between the local VABB-SHW database and OpenAlex given the limited availability of DOIs in our local database.

Introduction

OpenAlex has come onto the stage of large indexing databases in late 2022, taking over the backlog of the discontinued Microsoft Academic, and promising an open and non-commercial alternative to indexing databases. Unlike the proprietary alternatives, OpenAlex data can be shared freely under a CC0 license, which enables bibliometricians to share data openly. In the context of initiatives such as the Barcelona Declaration for Open Research Information (Kramer et al. 2024), this is a promising development for the field of bibliometrics. So far, OpenAlex has been used by several major institutions. The French Sorbonne university announced in 2023 that they would unsubscribe from Clarivate-owned Web of Science and opt for a partnership with OpenAlex[1]. Notably the latest Leiden Ranking, published by CWTS has added an open version using OpenAlex as a data source[2]. Another source for enthusiasm regarding OpenAlex is its promise to be both open and comprehensive. The OpenAlex website states: “We strive to be as comprehensive and inclusive as possible, especially for works in other languages and the Global South”[3].

[1] https://www.sorbonne-universite.fr/en/news/sorbonne-university-unsubscribes-web-science
[2] https://open.leidenranking.com/
[3] https://help.openalex.org/hc/en-us/articles/24396686889751-About-us

Insufficient coverage of non-English publications and insufficient coverage of the social sciences and humanities (SSH) is a researched limitation of the big international indexing databases (Mongeon & Paul-Hus, 2016; Kulczycki et al. 2018). This has a particularly significant effect on the representation of the SSH since authors from the SSH still publish more frequently in local, non-English language publication channels and books (Kulczycki et al. 2020; Giménez-Toledo 2020). Our goal is to examine to what extent OpenAlex covers SSH publications, including non-English language publications by using the comprehensive bibliographic database VABB-SHW (henceforth VABB) which includes all publications (co-) authored by researchers associated with SSH departments of Flemish universities. So far, around half of peer-reviewed records in the VABB database are covered by the Web of Science. However, Dutch-language publications are less likely to be covered by the Web of Science (only 1.5 percent of WoS-covered publications are classified as Dutch-language in VABB).

OpenAlex can be used as an open bibliometric data source, but for the SSH it is particularly important to track its coverage of diverse publication types and languages other than English. Much of the research on the coverage and metadata of OpenAlex is quite new, and not all has appeared in journal publication form by the time of writing. Researchers have investigated the reference coverage of OpenAlex, Web of Science and Scopus (Culbert et al., 2024) and found that OpenAlex performs similarly to the Web of Science and Scopus in terms of source reference coverage (an important difference is that OpenAlex does not include references to non-source items). Delgado-Quirós and Ortega found that while OpenAlex coverage is high, the source has a low completeness for bibliographic information (pages, issue, volume) (Delgado-Quirós & Ortega, 2024). In recent conference contributions, the coverage and metadata of African publications in OpenAlex, Scopus and Web of Science was investigated (Alonso-Álvarez & van Eck, 2024). Results show that OpenAlex outperforms Scopus and Web of Science in terms of coverage, and some metadata fields (notably ORCID) while underperforming in others. Another contribution has matched OpenAlex with the Norwegian Cristin database and found that OpenAlex covers almost all of the publications that have a DOI in the Cristin database (Armitage and Seland 2024). Maddi et al. (2024) have investigated coverage of Open Access journals in OpenAlex, Scopus and Web of Science and found that OpenAlex offers a comparatively more inclusive coverage of world regions and more balanced coverage of disciplines, with in particular a better representation of the social sciences. Researchers have also looked into the suitability of OpenAlex for bibliometric studies through a comparison with Scopus and concluded that analyses based on the Scopus master list can reliably be repeated with OpenAlex data, but also pointed to some areas of concern, including the completeness and accuracy of metadata, such as the language field (Alperin et al. 2024). Additionally, for the records indexed in the database, several concerns regarding data quality have been discussed. Zhang et al. (2024) concludes that institutional information is missing more frequently than in Web of Science. As mentioned, Delgado-Quirós and Ortega (2024) find that bibliographic information is frequently missing for OpenAlex records. Céspedes et al. (2024) determine that for 14.7% of papers in OpenAlex the

language declared on the platform is incorrect. Jiao et al. (2023) have found inconsistencies in the reporting of document types, with OpenAlex reporting all data articles as regular research articles. As OpenAlex uses data from the previously discontinued Microsoft Academic Graph (MAG), it initially inherited some of the properties of this earlier bibliographic service (Scheidsteger & Haunschild, 2023). OpenAlex lists as its main data sources MAG and Crossref but also sources such as Pubmed and arXiv, and adding additional metadata from ORCID, Unpaywall, ROR and others[4]. However, it is important to keep in mind that OpenAlex is evolving and improving quickly, which means that some issues reported in earlier studies may already be fixed by now.

In this study, we use a comprehensive database as a benchmark, which allows us not only to analyse how coverage compares to the Web of Science, but also to get a full overview of which publications are well-covered by OpenAlex and which publications are missing. The process of matching the regional database with OpenAlex through DOI, ISSN, title and author names is also an important element in this effort. We hope that this analysis may prove useful to researchers planning to use OpenAlex for bibliometric research that includes the SSH and non-English language sources in particular. Additionally, we also hope that this might be of interest to the community around OpenAlex which is working towards improving the database.

In this study we will focus on two main aspects. One is the data matching between the local bibliographic database and OpenAlex. We report on the number of records we were able to match with a record from OpenAlex by three different matching methods. We then report on the number of publications from VABB we were able to find in OpenAlex, the characteristics of those publications and the metadata coverage for those publications. Specifically, we are interested in the following:

1. How many records from VABB (2013-2022) can we find in OpenAlex with different matching strategies?
2. What are the characteristics of the publications we could find/could not find in OpenAlex?
    a. In terms of indexation in the Web of Science and peer review status
    b. In terms of language
    c. In terms of publication type
3. What is the metadata coverage of VABB publications in OpenAlex?
    a. Inclusion of reference/citation information
    b. Completeness of affiliation information

Data

As mentioned, we use the Flemish bibliographic database VABB to compare coverage of publications in OpenAlex. The VABB database is created and

[4] https://help.openalex.org/hc/en-us/articles/24397285563671-About-the-data

maintained as part of the Flemish performance-based research funding system. Part of the publications that are in VABB are also indexed in the Web of Science (37.7 %). A second part is not indexed in the Web of Science but is published in publication channels approved by an Authoritative Panel (GP) (32.9%). These publications are considered to be peer-reviewed. A third group of publications consists of publications that were not approved for various reasons (29.3%). We split this group into: publications that were not approved because of formal criteria (missing ISSN/ISBN, missing page info or under 4 pages long) and publications that were not included because they are not considered peer-reviewed by the GP. Figure 1 gives an overview of the peer-reviewed and non-peer-reviewed parts of the dataset. For this study, we use VABB records published between 2013 and 2022, including non-peer-reviewed publications. In the data cleaning process, we removed publications that were not considered as part of the peer-reviewed publications in VABB because they were of the wrong discipline (non-SSH), these were 503 publications. In total, this leaves us with a dataset of 146,680 publications to be matched with OpenAlex. The latest version of the peer-reviewed records in VABB can be accessed online (Aspeslagh et al., 2024).

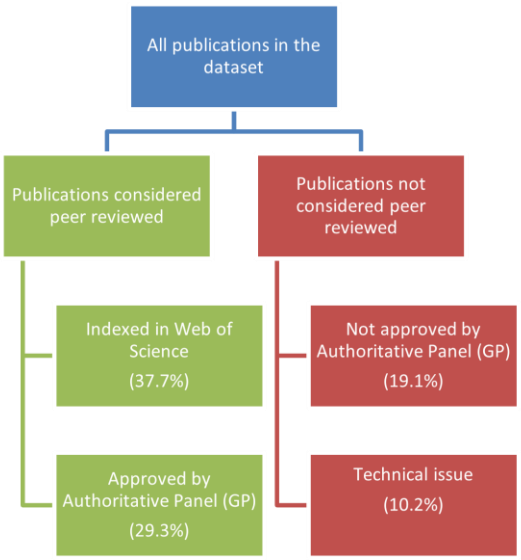


Figure 1. Schema of publications in VABB database.

VABB records belong to one of the following categories: journal article, conference proceedings paper, edited book, book chapter and monograph. Figure 2 gives an overview of the number of records in each of the publication types in VABB.

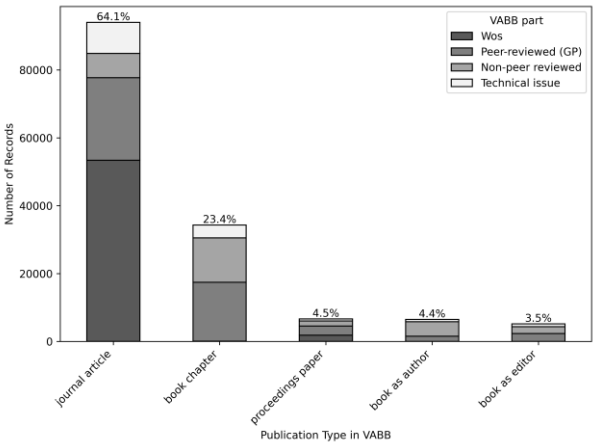


Figure 2. Number of VABB publications per publication type.

A majority (64%) of publications belong to the category journal article. The number of books has been decreasing over the years, but remains an important publication type for SSH although it is not covered well by the Web of Science.
We are using the OpenAlex snapshot of October 2024 (hosted by the Insyspo project). The records could be accessed through Google BigQuery.

Matching procedure

We have adopted a three-step search strategy for identifying VABB publications in OpenAlex (Figure 3). The first step is a matching based on DOI (digital object identifiers). The second step is a matching based on exact title, year (allowing for 1year difference) and at least one author. We chose to allow the publication year to be higher or lower to allow for variations related to preprints and online early access. A third step is a matching based on ISSN, year and author followed by a fuzzy title match. The fuzzy title matching uses the ratio Levenshtein distance. A ratio of above 0.80 is considered a match.

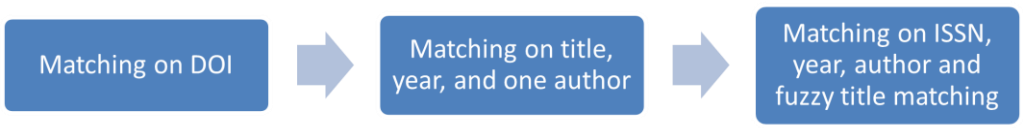


Figure 3. Overview of the matching procedure.

Results

Overview of the number of records matched with the three search steps

We were able to match 74,021 records from VABB to a record in OpenAlex, this is slightly over 50%. Most of the publications could be matched through DOI. Including

the other two search steps yields more incremental gains. Table 1 gives an overview of how many publications can be found with each of the steps.

Table 1. The number of records found with each of the matching steps and the percentage of total publications in VABB and the number of records added by including the step.

Search step	*Number of records found in OpenAlex*	*Number of records added by including the step*
Step 1: DOI	65,921 (44.94%)	65,921
Step 2: Exact title	34,636 (23.61%)	6,103
Step 3: Fuzzy title	53,028 (36.15%)	1,997

In total, 67,698 records in the VABB-SHW have a DOI identifier, which is 46.2 percent of the records. Matching on DOI yielded 66,014 matches in OpenAlex which means that 97.5 percent of records with a DOI could be found in OpenAlex. However, there are a few records for which the same DOI was associated with multiple records in VABB. This is the case in particular for book chapters where the DOI listed in VABB refers to the whole book rather than the individual chapter. We excluded these book chapters with the same DOI from the DOI results. In addition, there were 4 DOIs that yielded multiple work_ids. Upon reviewing, we found that one was a mistake in OpenAlex, on an erratum and two cases were preprints, these were excluded as well. With these cleaning steps we arrive at a final set of 65,921 records matched through DOI.

With 97.5 percent of DOI's matched in OpenAlex, matching with DOI has very good results. A recent conference contribution matching academic publications from the Norwegian Cristin database to OpenAlex yielded coverage of 99% for academic works and 97% for the other works (Armitage and Seland 2024). Figure 4 shows the annual number of records with and without DOI in VABB over the time period.

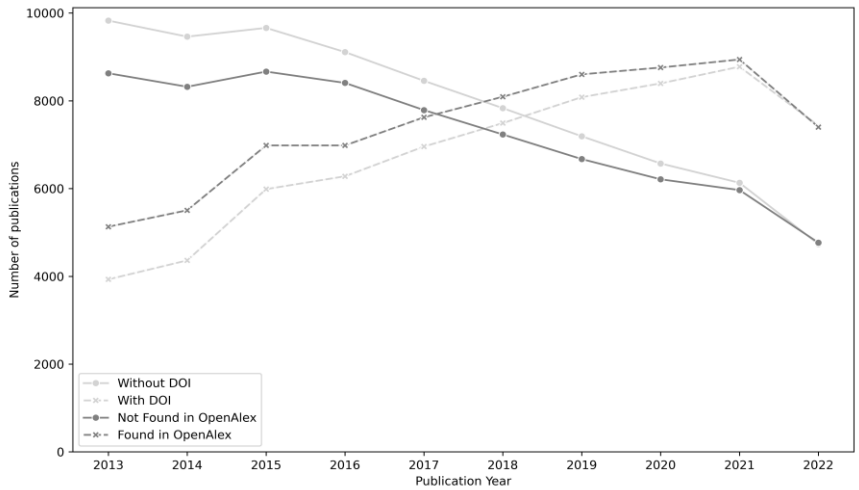


Figure 4. Evolution of the number of records with and without DOI in VABB (2012-2022).

The number of publications with DOI is increasing, while the number of publications without DOI is decreasing. Considering how well publications with DOI are covered in OpenAlex, we expect that the number of publications matched with OpenAlex will increase as more publications are issued a DOI. The increasing availability of DOIs for records in VABB largely tracks the increasing visibility of VABB records in OpenAlex.

The second step was a search for publications for which we did not find a matching DOI. This step consisted of a matching by title, publication year and at least one author. As mentioned above, we allowed the publication year to differ by one. We found that in some cases, the second step found multiple work-id's. In case of multiple work-id's we gave preference to the matching based on DOI for the final dataset as these are more likely to refer to the final publication. There is a significant overlap between publications found with DOI matching and the exact title matching.

A third step included a matching by publication year (again allowing a one year difference) and one author as in the previous step. Additionally, we matched on ISSN followed by a fuzzy title matching (using Levenshtein distance ratio of above 0.80). Evidently this search strategy only yields results for records with an ISSN (typically journal publications). All publications that are found in step 2 and that have an ISSN can also be found with the fuzzy title matching. Fuzzy title matching is more computationally intensive and therefore only an option as a 'last resort'. The number of additional records found with the fuzzy matching is limited (1,997).

Figure 5 shows that there is significant overlap in the results obtained with the three search steps, with DOI-based matching yielding the largest number of unique matches.

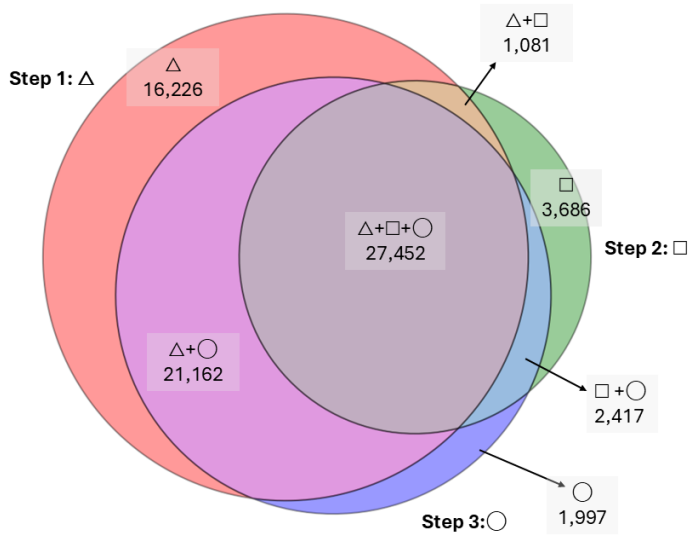


Figure 5. Overlap between records found in each of the search steps.

Characteristics of publications matched with OpenAlex

In table 2, we show the breakdown of coverage in OpenAlex for the publications that are also in the Web of Science, publications that are approved by the Authoritative Panel (considered peer-reviewed) and publications that are not peer-reviewed or do not count in the Flemish PRFS for technical reasons. We do not expect a high proportion of publications that are not considered peer-reviewed to be found in OpenAlex, as these may include grey literature and publications aimed at a broader audience, but we are including them for the sake of completeness. The breakdown shows that most of the publications from our database that are indexed in the Web of Science are also present in OpenAlex. This aligns with findings from previous studies on OpenAlex that have indicated that it provides good coverage for publications indexed in the Web of Science or Scopus (Alperin et al. 2024; Culbert et al. 2024). For publications that are not indexed in the Web of Science, the coverage is lower. Peer-reviewed publications that are not covered in the Web of Science, have a coverage of about 37 percent in OpenAlex.

Table 2. Number and percentage of publications found in OpenAlex according to the different parts of the VABB database (publications indexed in the Web of Science, other peer-reviewed publications (GP), non-peer-reviewed publications and publications not included for technical reasons).

Part of VABB	*Found in OpenAlex*
Indexed in WOS	52,315 (94.51%)
Other peer-reviewed (GP)	17,954 (37.15%)
Non-peer-reviewed	2,336 (8.35%)
Technical issue	1,416 (9.43%)
Total	74,021 (50.46%)

Considering the coverage per publication type (Table 3), we observe that journal articles are the most comprehensively represented, while only a small proportion of book publications are retrieved. This could be related to our methodology for the retrieval of the information from OpenAlex. It is possible that we are missing book publications because many book publications do not have DOIs and we were unable to conduct searches based on ISBN. Nevertheless, we can assume that coverage is better for journal articles, especially for journal articles in internationally visible English-language journals.

Table 3. Overview of publications found in OpenAlex. Breakdown by publication type in VABB.

Publication type	*Found in OpenAlex*
Journal article	63,560 (67.6%)
Book chapter	6,117 (17.83%)
Proceedings paper	3,124 (46.89%)

Book as editor	629 (12.12%)
Book as author (monograph)	591 (9.1%)
Total	74,021 (50.46%)

Coverage of non-English language sources is an ongoing concern for the social sciences and humanities. The multilingual nature of the VABB database allows us to investigate the coverage in OpenAlex for sources in languages other than English, which is of particular importance as it would constitute an advantage over other international data sources. Table 4 shows the coverage of sources in the most frequent publication languages in VABB. English publications are covered best, whereas Dutch sources are covered only about 8%. This suggests that OpenAlex does not cover Dutch language VABB publications very well.

Table 4. Overview of publications found in OpenAlex. Breakdown by language.

Language	*Found in OpenAlex*
English	68,819 (70.91%)
Dutch	3,087 (8.05%)
French	911 (15.94%)
other	549 (22.07%)
Spanish	356 (28.03%)
German	299 (16.35%)
Total	74,021 (50.46%)

This is of course partly related to the more limited DOI coverage for Dutch-language publications in general and the relatively higher share of book publications (book chapters, monographs and edited volumes) in Dutch language publications. Only 1,452 out of 38,334 Dutch language publications have a DOI associated with them in the VABB database.

The availability of references and affiliation information

Apart from coverage in OpenAlex, we are also interested in the availability of metadata. For bibliometric studies, the availability of metadata is of crucial importance. A quick note on the way in which OpenAlex deals with records is warranted here. OpenAlex is envisioned as a graph connecting different entities. Each of the different entities in the graph is accorded a unique identifier. There are works, authors, venues and institutions. These entities are connected to each other. OpenAlex does not record references to 'non-source' items. All references recorded also refer to a work entity in OpenAlex. In terms of metadata about institutions, OpenAlex assigns a ROR identifier to all institutions. This is a useful addition because it makes it easier to link the institutions to other datasets. OpenAlex also attaches considerable importance to ORCIDs. In previous studies it has been noted that OpenAlex makes more ORCIDs available than other bibliographic sources (Alonso-Alvarez and Van Eck 2024; Culbert et al. 2024). We study two aspects of metadata coverage: references and affiliation information. In terms of the coverage

of references we look at the number of publications that have at least one reference and at the median number of references per publication.
In total, 63,518 publications matched with OpenAlex include at least one reference, this is 86 percent of records. In table 5 we show the inclusion of references broken down by publications that are also covered in the Web of Science, publications that are peer-reviewed, publications that are non-peer-reviewed and publications that are not included because of technical issues. While reference coverage is high for publications that are also indexed in the Web of Science, there are a large number of publications with zero references for the other parts of the database. For the non-peer-reviewed publications and publications excluded for technical reasons, this may be a reflection of the nature of those publications, which may include grey literature, short reviews and editorial material. For the peer reviewed publications approved by the GP, the number of publications may be due in part to the publishers not providing access to the reference information, making it harder for references to be included in Open Alex. In terms of the median number of references per publication, we note the relatively high values for publications that are also indexed in the Web of Science.

Table 5. Number and percentage of publications in Open Alex that include at least one reference. Median number of references for records that have at least one reference.

VABB part	*Includes references in OpenAlex*	*Median number of references (for records with references in OpenAlex)*
Web of Science	50,169 (95.9%)	44
Other peer-reviewed (GP)	11,945 (66.53%)	27
Non-peer reviewed	847 (36.26%)	19
Technical issue	557 (39.34%)	12
Total	63,518 (85.81%)	41

The affiliation fields gathered from OpenAlex are the following: raw affiliation string, institution name, institution id, ROR identifier, country code (of the institution) and ORCID (of the author). As mentioned, OpenAlex assigns ROR identifiers to all affiliation instances. Affiliation information is completely missing for 6,432 publications (or 8.7 percent of records). For the other publications, there is at least some affiliation information present. We show the number of complete or missing fields per author in table 6.

Table 6. Number and percentage of publications in OpenAlex that have missing affiliation information.

Data field	*# Publications missing entries (total)*	*Missing entries – WoS part*	*Missing entries – peer-reviewed (GP) part*	*Missing entries – non-peer-reviewed part*	*Missing entries – technical issue*
ORCID	36,553 (49.4%)	25,386 (48.53%)	9,227 (51.39%)	1,200 (51.37%)	740 (52.26%)
Country	6,500 (8.8%)	1,973 (3.77%)	3,954 (22.02%)	320 (13.70%)	253 (17.87%)
Institution ID	6,432 (8.7%)	1,915 (3.66%)	3,946 (21.98%)	318 (13.61%)	253 (17.87%)
ROR ID	6,432 (8.7%)	1,915 (3.66%)	3,946 (21.98 %)	318 (13.61%)	253 (17.87%)
Institution name	6,432 (8.7%)	1,915 (3.66%)	3,946 (21.98%)	218 (13.61%)	253 (17.87%)

OpenAlex includes ORCIDs, although ORCID identifiers are not available for all authors. The reason for this is twofold. Not all researchers have ORCID profiles, and it is not always straightforward to link ORCIDs to researchers. Furthermore, OpenAlex links each institution to a ROR ID (which is why ROR IDs are available for most affiliation instances). However, it is not clear whether each of these links are accurate. Breaking down by peer-review status and indexation in the Web of Science, we can see that affiliation information is more available for publications that are also indexed in the Web of Science, and more likely to be missing for publications that are not. This is in line with other studies on metadata completeness in OpenAlex. Metadata is more available for journal articles and less for books and other publication types. These numbers do not give an indication of the quality of the metadata, which relies in large part on the performance of the disambiguation algorithms used by OpenAlex that connect authors to ORCID profiles and affiliation information to ROR identifiers.

Discussion

From the records found in OpenAlex we can gather that OpenAlex does include additional publications that are part of the VABB database but not covered in the Web of Science, but does not come close to covering all peer-reviewed publications in VABB. More specifically, Dutch language publications are not covered well and non-journal articles are also not covered well. There are some reasons for why this might be the case. The most successful way in which we were able to match publications across databases was through DOI. OpenAlex covers records with DOI quite well. This is probably due to the way in which records are added to the database. Crossref is one of the main sources of OpenAlex and is also one of the main DOI

registration agencies[5]. Records with DOI are more easily traceable and identifiable online. However, many publications do not have a DOI. In particular, books are frequently not assigned a DOI and many (local) journals similarly do not regularly assign DOIs. This is due to several reasons, including the fact that registering a DOI is not free of charge. A recent conference contribution of the coverage of publications from the CRISTIN database came to similar conclusions with regards to the inclusion of books and publications without DOI (Armitage and Seland 2024). This is important to keep in mind as studies may rely solely on DOI to match with OpenAlex.
In terms of the coverage of publications that are also covered in the Web of Science, OpenAlex covers a large majority of publications. The publications that were not found in this way could be due to several reasons, including incomplete or inaccurate data in one of the sources (missing DOI in VABB, title variations, etc). The coverage of records that are not considered peer-reviewed in VABB is lower, which is understandable considering OpenAlex's focus on research publications.
An overview of the metadata covered in OpenAlex gives us insight into its potential usefulness to enrich our local database and use for the purposes of bibliometric research. The VABB database does not include reference information, meaning that a citation analysis of the Flemish SSH needs to rely on additional data sources. While OpenAlex does not offer broad coverage of non-English language SSH literature, it offers more comprehensive coverage than the Web of Science.

Limitations

First, we have to note that our search strategies do not exhaust all of the possible ways in which records could be matched with OpenAlex. Alternative approaches could focus on ISSN coverage as a proxy, allow for errors in author names etc. We have tried here to use an approach that could potentially be replicated with other publication databases.
Second, it is possible that there are publications in VABB that have a DOI that is not in our database. The VABB records as many DOIs as possible, and DOIs are frequently added as part of the data enrichment process, but universities are not required to add DOIs to publications they submit to VABB, which means that DOI coverage in VABB is not complete.
Third, we should note that OpenAlex is changing rapidly. We have used a snapshot of October 2024, but it is possible that by the time of the conference, the results of this exercise may differ.
Fourth, our results with regards to metadata only include whether or not a field was available for a particular record. Our results do not provide evidence to the quality or accuracy of the metadata included. Additional research could look further into the quality of references and affiliation information.

[5] https://www.crossref.org/

Conclusion

We have matched records from the local bibliographic database VABB with OpenAlex and reported on the results of matching with several search strategies and the coverage of OpenAlex across language, publication type, indexation in Web of Science and peer review status. Our main conclusions are that OpenAlex provides good coverage of publications with DOI, which means that it covers the parts of the local database that have a DOI (mainly journal articles and publications in English). This also means that coverage for books, and publications in languages other than English is low. In terms of metadata, OpenAlex provides most metadata for records that are also found in the Web of Science, but also includes metadata for many of the records that are not included in the Web of Science.

From the perspective of open data, the high number of references available in OpenAlex is an exciting possibility to use open and non-proprietary data.

We think the results of this research could be of interest to the bibliometric community, the community around OpenAlex and also local publishers who would like to increase the international visibility of their scholarly publications in OpenAlex. International bibliographic databases usually perform worse for the SSH and for publications in languages other than English, which poses difficulties for bibliometricians interested in those fields. While we could retrieve more publications from the comprehensive regional database in OpenAlex than in Web of Science, there is still a large number of publications that were not found in OpenAlex. While some of these discrepancies could be explained by the obscurity of the material (publications that are not strictly scholarly), many of the publications are peer-reviewed scholarly materials. Improvements to OpenAlex could include making searches based on ISBN easier and attempting to include more book publications. For local publishers, we think these results show that registering DOIs increases visibility in OpenAlex. Coverage of non-English language sources will improve if more records are assigned a DOI. Alternatively, adding the records from VABB (and by extension other national bibliographic databases) to OpenAlex could be an interesting way forward. Adding VABB data to OpenAlex would increase the visibility of the Flemish SSH. For bibliometricians, our results indicate that caution is warranted when performing bibliometric studies focusing on the SSH with OpenAlex. Coverage of non-English language sources and book publications is still relatively low, even if it is higher than for alternative sources (notably the Web of Science). OpenAlex is, at this point, a valuable source to enrich the local database, but it is not at the level of replacing it.

Acknowledgements

We used a version of OpenAlex hosted by the Insyspo project and are very grateful for the access to this resource.

References

Alonso-Álvarez, P., & van Eck, N. J. (2024). *Coverage and metadata availability of African publications in OpenAlex: A comparative analysis*. Zenodo. https://doi.org/10.5281/zenodo.14006425

Alperin, J. P., Jason Portenoy, Demes, K., Larivière, V., & Haustein, S. (2024). *An analysis of the suitability of OpenAlex for bibliometric analyses* (arXiv:2404.17663). arXiv, 26 april 2024. https://doi.org/10.48550/arXiv.2404.17663

Armitage, C., & Seland E. H. (2024). *NWB2024: Will OpenAlex Solve Our Problems? A Coverage and Metadata Comparison with Cristin* [Conference presentation], 16 december 2024. https://doi.org/10.6084/m9.figshare.28033397.v1.

Aspeslagh, P., Guns, R., & Engels, T. C. E. (2024). *VABB-SHW: Dataset of Flemish Academic Bibliography for the Social Sciences and Humanities (edition 14)* [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.14214806

Céspedes, L., Kozlowski, D., Pradier, C., Holmberg Sainte-Marie, M. Shokida, N. S., Benz, P., et al. (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness'. *Journal of the Association for Information Science and Technology*, 1-12. https://doi.org/10.1002/asi.24979.

Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). *Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus* (arXiv:2401.16359). arXiv. https://doi.org/10.48550/arXiv.2401.16359

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. *Quantitative Science Studies*, 5(1), 31–49. https://doi.org/10.1162/qss_a_00286

Giménez-Toledo, E. (2020). Why books are important in the scholarly communication system in social sciences and humanities. *Scholarly Assessment Reports*, 2(1), 6. https://doi.org/10.29024/sar.14.

Jiao, C., Li, K., & Fang, Z. (2023). How are exclusively data journals indexed in major scholarly databases? An examination of four databases'. *Scientific Data*. 10(1), 737. https://doi.org/10.1038/s41597-023-02625-x.

Kramer, B. , Neylon, C and Waltman, L. (2024). *Barcelona Declaration on Open Research Information*, Zenodo. https://zenodo.org/records/10958522.

Kulczycki, E., Guns, R., Pölönen, J., Engels, T.C.E, Rozkosz, E., Zuccala, A., Bruun, K. et al. (2020) Multilingual Publishing in the Social Sciences and Humanities: A Seven-Country European Study. *Journal of the Association for Information Science and Technology* 71(11): 1371-1385. https://doi.org/10.1002/asi.24336.

Maddi, A., Maisonobe, M., & Zeghmouri, C.B. (2024) *Geographical and Disciplinary Coverage of Open Access Journals: OpenAlex, Scopus and WoS*. OSF. https://doi.org/10.31235/osf.io/8wa4q.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, 106(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Scheidsteger, T., & Haunschild, R. (2023). Which of the metadata with relevance for bibliometrics are the same and which are different when switching from Microsoft Academic Graph to OpenAlex? *Profesional de La Información*, 32(2), Article 2. https://doi.org/10.3145/epi.2023.mar.09

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, Y. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics* 129. 5869-5891 https://doi.org/10.1007/s11192-023-04923-y

https://doi.org/10.51408/issi2025_096

Multi-Disciplinal, Large Scale Mentorship Dataset and Demographics

Chiaki Miura [1], Yoshiyasu Watanabe[2], Tsubasa Sakammoto[3], Hiroshi Hashizume[4]

[1]*1t.miura@gnt.place*, [2]*watanabe.yoshiyasu.2m@kyoto-u.ac.jp*,
[3]*sakamoto.tsubasa.5n@kyoto-u.ac.jp*, [4]*hashizume.hiroshi.8z@kyoto-u.ac.jp*
Department of Engineering, The University of Tokyo, Bunkyo, Tokyo (Japan)
Office of Research Acceleration, Kyoto University, Kyoto (Japan)

Abstract

Academic genealogy depicts a relationship network of mentor-trainees, which embraces the rich history of knowledge flow through discipline. We matched over 800 thousand researchers to the world's largest academic genealogy database with the open/libre bibliographic database of over 200 million research works through author names, works, and institutions. This allows scientometricians and higher education strategists to conduct comprehensive analyses of researcher mobility, training, institutional bias, and success. The paper also provides the complete descriptive statistics and propensity of the Academic Family Tree dataset.

Introduction

Mentoring in academia is not only an act of learning, but a profound mecha nism of knowledge transmission. As Zuckerman (1977)[9] and others have ob served, academic disciplines are mediated by formal and informal norms, many of which are implicitly transmitted through interactions between young scien tists and their mentors. Because such interactions are important moments of tacit knowledge exchange across academic fields, the genealogy of mentors and trainees provides a quantitative framework for exploring these relationships and their broader impact on academic ecosystems (7; 4; 1) .

Existing research has emphasized the importance of mentorship in academic career development. Studies from various fields have shown that mentors with high mentorship fecundity, who produce many trainees, increase their scien tific legacy through the success of their students. For example, Sugimoto et al. (2011) (8) demonstrated that the field of expertise of a supervisor directly affects the interdisciplinary nature of a student's dissertation, emphasizing the role of mentorship in the formation of intellectual paradigms. Tol (2024) [8], who recently used the Academic Family Tree dataset to integrate the academic lineage of Nobel Prize winners, also points out that the lineage of academic men tors not only promotes excellence, but also leads to close intellectual networks. These insights highlight the depth of the structural impact of mentorship in cultivating groups of elite scientists.

The "Academic Family Tree", pioneered by David and Hayden 2012 (3) for its antecedent known as"Neurotree", provides a unique opportunity to analyze the relationship between mentors and their trainees on an unprecedented scale. This dataset contains bibliographic record of over 876 thousand scientists and 1.8 million

liaisons, and it is possible to understand how mentoring relationships affect scientific productivity and success. Unlike traditional case studies with limited generalizability, this large-scale dataset enables rigorous statistical analysis across disciplines and institutions.

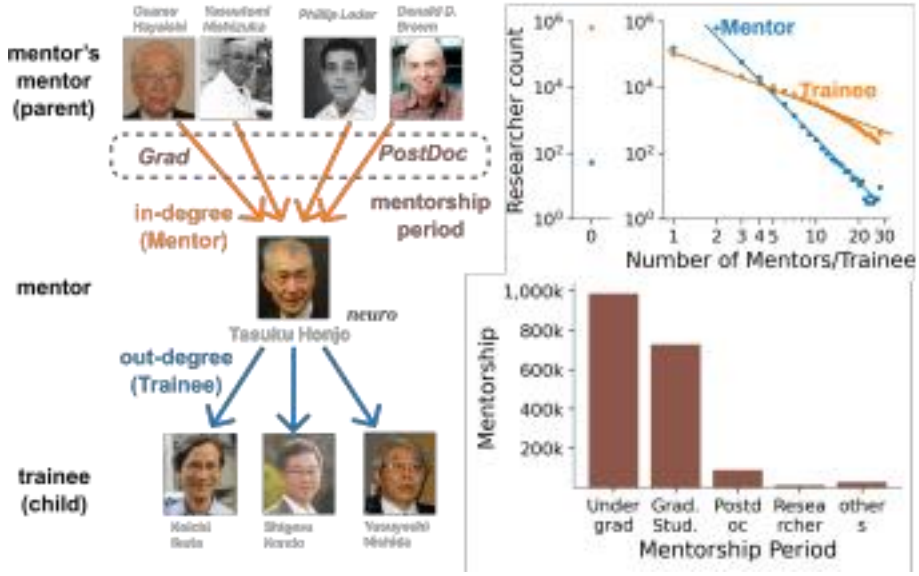


Figure 1: Descriptive statistics of academic family tree dataset.

Ke et al. (2022) (6) combined datasets about mentorship with the Microsoft Academic Graph (MAG) to identify patterns of mentor effectiveness and demographic differences. Building on the findings of these previous attempts, this paper proposes to integrate the Academic Family Tree and OpenAlex – a fully open bibliographic database developed as the successor to the MAG – to present a systematically managed database that allows more scalable analysis of academic genealogy. (2)

Academic family tree is the world's largest human-annotated academic genealogy database. It is later expanded largely by the American dissertation repository (ProQuest). Most of the mentorship relationships registered are mentorships during undergraduate education and graduate student training. It is remarkable considering that the number of graduate students is almost the same as that of postdocs in the same cohort (16,000 in 2000 to 13,000 in three years after that, four major areas aggregated) (5).

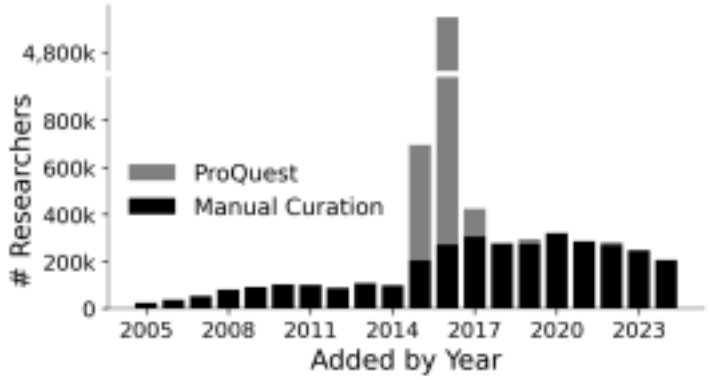


Figure 2. Added by Year.

One of the main questions underlying this research is as follows: What characteristics of the mentor-trainee relationship predict the academic success of the trainee? While previous research suggests that successful mentor is most likely train a successful trainee, the underlying mechanisms are still unclear. Is success primarily a function of intellectual compatibility when the mentor's and trainee's areas of study coincide? Is success due to the mentor's ability to secure access to influential networks and resources? What are the specific mechanisms by which tacit knowledge, such as awareness of grant opportunities or potential collaborators, is transferred from mentor to trainee? Or can these pathways give rise to biases, and how can identifying them help overcome existing barriers to equitable academic advancement? This paper will provide a solid foundation for a more nuanced understanding of the impact of mentorship on academic careers and guide the development of policies and initiatives to support the next generation of scholars.

Table 1. Researcher entity and coverage of Academic Family Tree Attribute Data Count.

Attribute	Data Count	Coverage (%)
First name	875,162	99.87
Middle name	448,272	51.15
Last name	874,974	99.84
Degrees	470,156	53.65
Location	876,147	99.98
Major Area	876,230	99.99
Areas	626,569	71.50
Award	866	0.10
h-index	379,356	43.29
ORCID ID	6,821	0.10
Semantic Scholar ID	379,356	43.28
Homepage	67,519	7.70
Added by	876,228	99.99
Date Added	876,330	100.00

Table 2. Relationship entity and coverage of Academic Family Tree Attribute Data Count Coverage (%).

Attribute	Data Count	Coverage (%)
Mentorship Period	1,841,686	100.00
Location[1]	1,841,103	99.97
Dissertation Title	520,870	28.28
ProQuest ID	438,576	23.81
Added by	1,840,868	99.96
Date Added	1,841,686	100.00

*Location is complemented by Location ID (locid)

OpenAlex

This study maps the authors in the two bibliographic databases and investigates the statistics and registration bias in the AFT dataset. Main source of OpenAlex authors are from the authorship in the works that are mainly retrieved from Crossref, and information about authors "comes from MAG, Crossref, PubMed, ORCID, and publisher websites." OpenAlex then disambiguates and aggregates author records based on how well authors with the same name share a tendency of their works. This algorithm allows us to incrementally aggregate differently written author names and is robust against spelling inconsistencies.

Method and Materials

We used the snapshot of Academic Family Tree (AFT) taken on Oct. 2024. Out of 876,304 researchers on the AFT dataset, 1,168 (0.13%) and 1,356 (0.15%) are missing first name and last name, respectively. We removed records whose first name and last name are both missing, which is equivalent to 920 researchers. We did a few more cleansing, and name and ID normalizations were done to get the best matching accuracy (see Supplementary 1). The whole procedure is depicted in Fig.3. Here, we took ORCID ID as a gold standard, which yields 6,766 matched researchers between the two databases (see supplementary 2). Among the rest, around half (51.2%) have a middle name. Coverage of other major columns are 100%, 98.2%, 53.6%, and 7.3% for major area, location, degree, and homepage, respectively [Table 1.].

As a preliminary result, here we propose a result from the sample of 10,000 AFT records. We conducted all the matching through OpneAlex API between Jan. 5. 2025 and Jan. 10. 2025. In the final version of this matching is done on OpenAlex full snapshot. We first took each author record in AFT dataset, and searched via OpenAlex API using the author's full name as a query.

Figure 3. Matching procedures.

Result and Discussion

AFT records is compared with OpenAlex author demographics, which reflect the widest possible researcher population who ever published any global report. Fig4 a. shows over- or under- representation of the countries author, namely the relative registration ratio of the county compared to the share of researchers in the world. The country of the author was inferred from the location of author's registered institution. US, UK, and French colonial institutions have higher registration rate than other countries, among other well represented developed countries in Europe. Similar disparity is between disciplines (fig4 b). Although the disproportionately high neuroscience representation is due to that the ser vice started in the discipline and accumulated most effort. Researchers from psychology, biochemistry have a higher registration rate than average, followed by nursing, medicine and immunology, which may reflect the disciplinal prox imity. Furthermore, the registered researchers are renowned researchers; they have higher mean impact and productivity, with median 2.3×10^2 and 3.2×10^4 times larger than the average, respectively(fig.4 d). Note that the both y axis for impact and productivity are log scaled. Surprisingly, nonetheless the registered researcher does not have significantly different academic age, which is consistent through all the cohort of the career start year (fig.4 c). Gender imbalance is slightly higher than global average

(fig.4 e). Expected ratio is calculated from the weighted mean of inferred degree of the registered researchers.

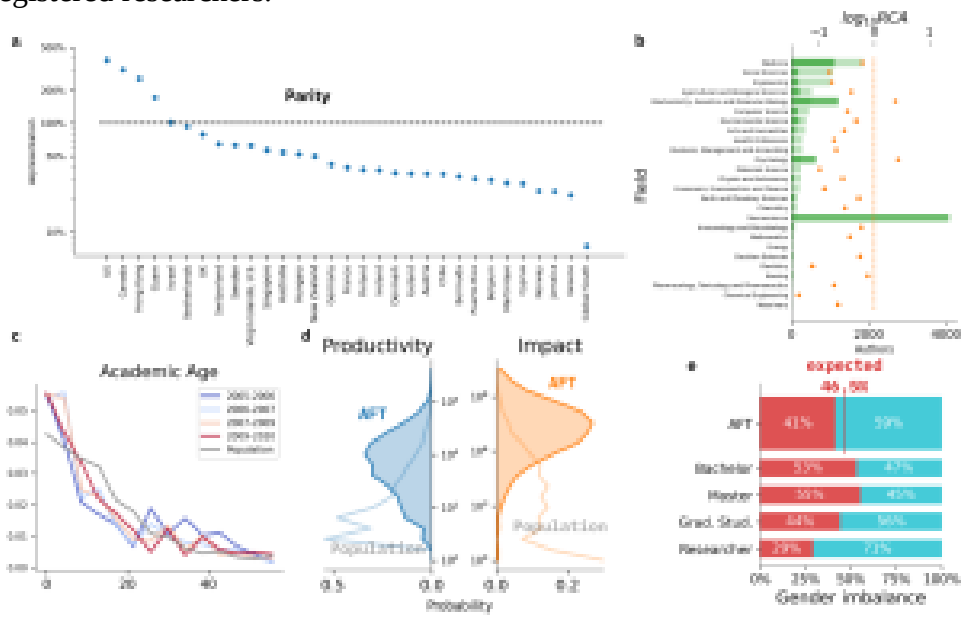


Figure 4. demographics of AFT. a) Top 30 most represented countries in AFT. b) Representation difference by fields. c) Academic age demographics, strati fied by the registration year cohort. d) Distribution of authors with a certain productivity and impact. e) Registered authors gender balance. Following bars shows the global imbalance by degree.

Academic Family Tree is a community-supported registration server that covers researchers and their mentor-trainee relationship from various background. Although it has some degree of registration bias, academic genealogy can yield a rich information on the knowledge flow, if dealt with an adequate calibrations.

Acknowledgements

Stephen David gave us a thorough instruction on academic family tree data format interpretation.

References

Katy B¨orner, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans. Skill discrepancies be tween research, education, and jobs reveal the critical need to supply soft skills for the data economy. 115(50):12630–12637.

Nicolas Carayol and Thuc Thi. Why do academic scientists engage in inter disciplinary research? 14(1):70–79. Publisher: Oxford University Press.

Stephen V. David and Benjamin Y. Hayden. Neurotree: A collaborative, graphical database of the academic genealogy of neuroscience. 7(10):e46608. Publisher: Public Library of Science.

James A. Evans. Industry collaboration, scientific sharing, and the dissem ination of knowledge. 40(5):757–791.

National Institutes of Health and others. Biomedical research workforce working group report.
Qing Ke, Lizhen Liang, Ying Ding, Stephen V. David, and Daniel E. Acuna. A dataset of mentorship in bioscience with semantic and demographic esti mations. 9(1):467. Publisher: Nature Publishing Group.
Katia Levecque, Frederik Anseel, Alain De Beuckelaer, Johan Van der Hey den, and Lydia Gisle. Work organization and mental health problems in PhD students. 46(4):868–879.
Cassidy R. Sugimoto, Chaoqun Ni, Terrell G. Russell, and Brenna Bychowski. Academic genealogy as an indicator of in terdisciplinarity: An examination of dissertation networks in library and information science. 62(9):1808–1828. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.21568.

Supplementary

Data Availability

Our data and code will available at our project repository. The matched ID and other datasets will be uploaded on Zenodo as well.

Data Normalization

Name normalization

Some of the records have non-English names, nicknames, and other supplementary names, all of which we could observe were parenthe sized. We store those records separately in"rawname oaid.csv". Punctuation marks and other special latin characters are not modified. One record on AFT has ORCID while both firsname and lastname are missing (pid=944562). We took this and matched to openalex. On the other hand, two records on AFT has middlename while both firsname and lastname are missing (pid=879367, 929462). As these records as unreliable, we ignored this record throughout the process.

ORCID

Some of the ORCIDs are recorded on AFT while it is not disclosed on ORCID as a public record, result in no match on OpenAlex.

Semantic Scholar ID (s2id)

AFT dataset does have an id column to store semantic scholar ids, which is the numerical string at the end of the URL in the semantic scholar profile page and can be retrieved via API. Semantic Scholar has 79 million author records (viewed on Jan. 18, 2025) which is comparable to openAlex (101 million). We did not use them to match authors because 1. ORCID is a nonproprietary while S2ID is not, 2. OpenAlex does not currently support semantic scholar id in their database.

https://doi.org/10.51408/issi2025_112

What Are We Missing? A Systematic Approach to Overlap Analyses of Local and Global Repositories

Simon Willemin[1], Gaël Bernard[2], Julian Dederke[3], Mahmoud Hemila[4], Michelle Koch[5]

[1] simon.willemin@library.ethz.ch, [3] julian.dederke@library.ethz.ch, [4] mahmoud.hemila@library.ethz.ch,
[5] koch_michelle@outlook.com
ETH Zurich, ETH Library, Rämistrasse 101, CH-8092 Zurich (Switzerland)

[2] gael.bernard@epfl.ch
École Polytechnique Fédérale de Lausanne (EPFL), Institutional Data and Analytics Unit, CH-1015 Lausanne (Switzerland)

Abstract

Local repositories, managed by institutions, often differ in coverage and metadata from the research output affiliated with the concerned institutions in global repositories such as OpenAlex, which aggregate records from numerous sources for broader visibility. This paper introduces a DOI Screening System that systematically identifies and explains mismatches between local and global repositories by classifying publications as local-only, matched, or global-only. The system applies predefined rules and allows identifying patterns such as misattributed affiliations, unrecognized DOI prefixes, incomplete metadata, or underrepresented publication types. Based on these patterns, one can derive 'curative' actions. We demonstrate the system's utility by comparing the repositories of EPFL and of ETH Zurich to OpenAlex, showing how subtle inconsistencies in identifiers and affiliations can account for many discrepancies. The system provides insights into how targeted interventions addressing the root causes of these discrepancies can be used to enhance coverage and reliability in both local and global repositories.

Introduction

The landscape of bibliometric data has expanded considerably in recent years, with numerous openly accessible repositories complementing established, subscription-based platforms. Several global repositories (i.e., bibliometric databases) such as OpenAlex and OpenAIRE now coexist alongside institutional or national repositories, each serving distinct yet complementary purposes. *Local repositories* for scholarly outputs give institutions control over their data, provide archival continuity, and capture the full breadth of their scientific production—features critical for accurate record-keeping and institutional sovereignty. Conversely, *global repositories* of scholarly metadata bolster discoverability, expand the global reach of publications, facilitate benchmarking across institutions, and influence university rankings or decision-making, and therefore command significant attention from research administrators. Both types of repositories play a major role in international and national initiatives such as Plan S, the European Open Science Cloud (EOSC) and the Swiss National Open Access Strategy to ensure that research output is findable and accessible.

However, there are still discrepancies between the research output available locally and globally. Moreover, there is no systematic method to clarify why certain publications appear in one repository but not the other, and there are few tools to quickly perform an "overlap analysis" that compares the extent of coverage between several data sources. Similar to the biblioverlap package (Vieira & Leta, 2024), we aim to offer a semi-automated approach that allows anyone to perform such analyses. While our current system focuses on the overlap between a local and a global repository, our approach not only aims to identify the gaps, but also seeks to uncover the underlying reasons for these discrepancies, enabling a better understanding of the factors contributing to the mismatches. Systematically identifying the reasons behind overlaps (or lack thereof) among repositories can lead to concrete curation actions—such as updating metadata or affiliations. We refer to this approach to overlap analysis as *curative* in the sense that it aims to identify gaps to ultimately improve the coverage and metadata quality in both local and global repositories. It therefore goes beyond *descriptive* approaches that merely aim to understand the logic behind the selection and indexing process of a repository.

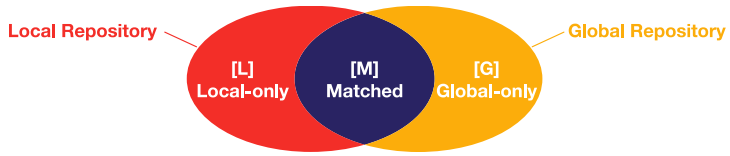


Figure 1. Sets representation for the overlap analysis.

As illustrated in Figure 1, a basic overlap analysis comparing local and global repositories typically categorizes publications into three groups: local-only, matched, and global-only. However, institutions need more than just these counts—they require practical explanations of why a publication is missing from one repository, whether due to issues like unregistered prefixes, incomplete metadata, or incorrect institutional attributions. Hence, we present a system that we call *DOI Screening System* and which is designed to help quickly gauge the overlap and the gap between a local and a global repository, as well as to deliver a list of indicators that can be used for curative purposes. The system takes a minimum set of information as input and automatically queries a global repository, classifies each publication into local-only, matched, or global-only, and applies automated rules to pinpoint the likely reasons for any discrepancies. This enables institutions to curate the related metadata by improving them or correcting institutional attributions where needed. Designed to be easily extended to additional repositories and new explanatory rules, the tool is available on GitHub, allowing for community-driven enhancements over time.
In the next section, we introduce the DOI Screening System and demonstrate its utility by comparing two local repositories to one global repository.

DOI Screening System

Description of the system

We present a *DOI Screening System* that automatically compares DOIs from a local and a global repository to identify which DOIs are missing from each source, as well as the reasons for these gaps. By classifying publications into local-only, matched, and global-only and then applying a set of predefined rules, the system provides insights that allow deriving actionable, "curative" steps to improve metadata accuracy and institutional coverage. The code is publicly available on GitHub (https://github.com/gaelbernard/DOI-screener), and can be extended to any repository or adapted with new rules as needed. Figure 2 shows an overview of the system.

Figure 2. Overview of the DOI Screening System.

Input. The system requires three minimal inputs: the ROR ID (Research Organization Registry Identifier), capturing the institution of interest, a year range specifying the temporal scope of the analysis, and a list of DOIs from the local repository, structured as a list of lists to accommodate multiple DOIs per publication. These inputs minimize setup complexity while still enabling a robust overlap analysis. A common source of error is incorrectly formatted DOIs, which may fail to be matched even after basic normalization. In addition, a limitation of the system is that it currently does not work for publications that do not have any DOIs.

Steps. Figure 2 illustrates the overall workflow, which consists of four main steps. First, the system queries the global repository using the specified ROR ID and chosen year range to collect all corresponding DOIs. Our system uses a deterministic approach to match DOIs, applying minimal text normalization (e.g., converting to lowercase, removing the "https://doi.org/" prefix). In its current iteration, the system uses only OpenAlex as a global repository (Priem et al., 2022), but it can readily be adapted to incorporate other data sources. Second, based on these global DOIs and the local repository's DOI list, it categorizes each publication into one of three sets: local-only (present locally but not globally within the expected time range and affiliation), matched (present locally and globally within the expected time range and affiliation), and global-only (present globally within the expected time range and affiliation, but not locally). Third, the system queries the global repository again—this time querying each unmatched local DOI instead of filtering by institutional affiliation or publication year. The DOIs retrieved during this third step are placed in the local-only category, that hence contains DOIs present locally but not globally or present globally without the expected time range or affiliation. Finally, the system applies a predefined set of rules (see Table 1) to diagnose why a DOI may not appear in both sources. Such a diagnose can be used to identify underrepresented output or inaccurate metadata and to target curation actions.

Table 1. List of rules implemented in the system.

Rule name	*Result from screening*	*DOI is present in Local list (L) or Global repo (G)*	*Description*
L-other	Unmatched	L	DOI from L that does not satisfy any other rule
L-prefix	Unmatched	L	DOI has a DOI-prefix that is significantly more frequently unmatched than matched (odds ratio)
L-time	Unmatched	L & G	DOI is outside the time range in G
L-inst	Unmatched	L & G	DOI is not affiliated with the institution in G
Matched	Matched	L & G	DOI is affiliated with institution and is within time range in G
G-prefix	Unmatched	G	DOI has a DOI-prefix that is significantly more frequently unmatched than matched (odds ratio)
G-type	Unmatched	G	DOI has a public. type that is significantly more frequently unmatched than matched (odds ratio)
G-authors	Unmatched	G	DOI has an author (identified through ORCID) that is affiliated with the institution in at least one of the matched DOIs in the same year
G-other	Unmatched	G	DOI from G that does not satisfy any other rule

These rules focus on issues such as misattributed affiliation (L-inst), out-of-range publication years (L-time), potential underrepresentation of certain sources identified through DOI prefixes (L-prefix / G-prefix), or of certain publication types in the local repository (G-type). Users can tailor or expand these rules to distinguish specific prefixes, to address unique metadata fields, or institution-specific patterns.

Output. The DOI Screening System provides two main outputs. The first is an *Overlap Bar Chart* visible in Figure 3 that shows how many publications fall under each rule, offering an immediate snapshot of the most common coverage or metadata issues. The second output is a detailed report that not only identifies which DOIs match each rule, but also provides additional information specific to each rule, such as the list of problematic prefixes. This enables librarians and research administrators to pinpoint and correct specific issues, such as updating metadata fields or resolving institutional attribution errors. Overall, this semi-automated approach offers both descriptive insights (quantifying the degree of overlap) and actionable items (pinpointing causes for mismatches) that enable curative measures, helping

institutions to curate data sources and to maintain robust, accurate bibliometric records. Although all rules are tested on each publication and appear in the detailed report, the order in which these rules are applied affects the distribution of publications in the bar chart output.

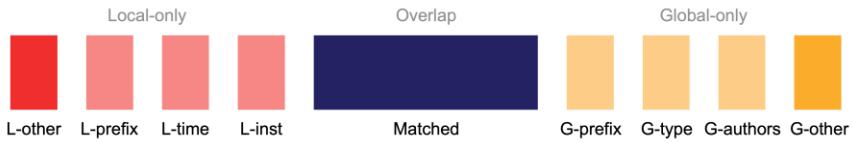


Figure 3. Bar chart representation for the overlap analysis with DOIs categorisation.

Two case studies

We applied the DOI screening system to the local repositories Infoscience at EPFL (ROR ID: https://ror.org/02s376052) and the Research Collection at ETH Zurich (ROR ID: https://ror.org/05a28rw58). OpenAlex served as the global repository, and the analysis covered the publication period from 2019 to 2023. For ETH Zurich, 46,579 publications were analyzed, with 9,518 (20.4%) not found in OpenAlex and 22,410 from OpenAlex not appearing in the local repository. At EPFL, 24,151 publications were analyzed, of which 5,369 (22.2%) did not appear in OpenAlex and 12,345 were found in OpenAlex but not in the local repository. The resulting Overlap Bar Charts are visible in Figure 4.

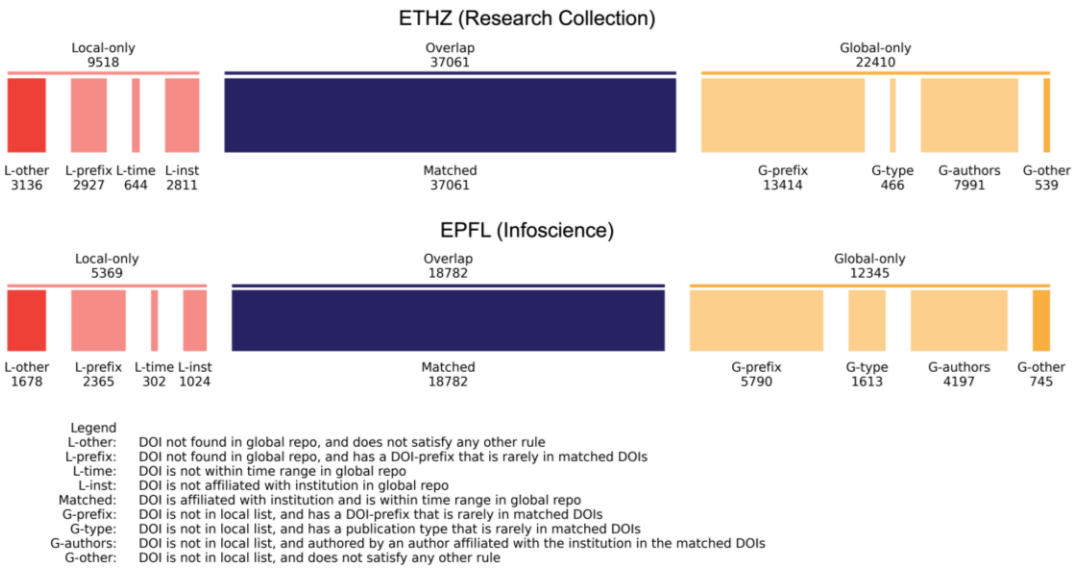


Figure 4. Visual output of the DOI Screening System based on OpenAlex, for DOI lists from local repositories of ETH Zurich and EPFL.

The system's predefined rules provide more detailed explanations for these mismatches that allow for informed decisions about the next curation steps. At EPFL, 44.0% (2,365) of local-only publications fell into the L-prefix category. Specifically, the system identified two prefixes, "10.5075" and "10.5281", that account for all

these mismatches: for example, “10.5075” appears in 2,280 local records but only 3 times in the global repository. Upon further investigation, we discovered that these prefixes are associated with EPFL theses, explaining why they are not automatically indexed in OpenAlex. Another 5.6% (302) were flagged as L-time, indicating that their publication dates fell outside the specified period and may require verification. In addition, 19.1% (1,024) were assigned L-inst, suggesting that these publications appear in OpenAlex but are not linked to EPFL, potentially requiring a curative action in the form of affiliation updates from the global repository. The remaining 31.3% (1,678) could not be explained by the current rules (L-other). Among global-only items, 46.9% (5,790) fell under G-prefix; for example, prefix “10.7910” appears 258 times in OpenAlex but never in the local repository, pointing to possible ingestion of new data sources locally. Another 13.1% (1,613) were categorized as G-type, as the system tagged peer-review, dataset, paratext, book-chapter, or preprint as underrepresented in the local repository. A further 34.0% (4,197) were flagged as G-authors, potentially indicating a missing research output in the local repository or a misattribution of affiliation in the global repository. The final 6.0% (745) were labeled G-other.

A parallel analysis at ETH Zurich revealed similar patterns, including publications not found in one source due to prefix or affiliation reasons, but with a notably lower percentage (2.0%) in the G-type category compared to 13.1% at EPFL. These results highlight how institutional policies or repository practices may influence coverage. Overall, the case studies demonstrate the value of the DOI Screening System in diagnosing coverage gaps, identifying metadata errors, and guiding targeted interventions to improve alignment between local and global repositories using a minimal set of input data.

Related Works

In this section we highlight some previous overlap analyses that are closest to our paper. Bologna et al. (2022) characterize studies on the coverage of global repositories including Web of Science, Scopus, Dimensions, Google Scholar and Microsoft Academic. Such analyses usually aim to determine the biases and provide an understanding of the selection processes at hand in global repositories (see also Delgado-Quirós, L. et al., 2024, Martín-Martín et al., 2021). They help researchers select the most appropriate data source and better evaluate the scope and limitations of the indicators they compute using such sources (Hug et al., 2017). Such analyses require strategies to match as many records as possible (see for instance Guerrero-Bote et al., 2021), in a context where global repositories are considered as relatively stable research objects.

Overlap analyses typically focus on comparing publications output or citations included in different repositories. In this study, we take a step further by categorizing unmatched publications according to a set of rules aiming to identify potential for improvements, an approach we characterize as *curative*. Descriptive and curative approaches should not be strictly contrasted, as recent papers focusing on open data sources illustrate. For example, Alperin et al. (2024) do not only explicitly compare OpenAlex and Scopus but also address critically the weaknesses of OpenAlex such

as accuracy and completeness of metadata. Hug and Brändle (2017) and Andreose et al. (2025) on the other hand explicitly compare a university's institutional bibliographic repository with Microsoft Academic or OpenCitations, respectively, as global repositories. This comes closest to our comparison of local and global repositories. With our curative approach, the goal is not primarily to describe the coverage and biases of the considered data sources regarding their suitability for research assessment. Instead, we explicitly aim to identify gaps and, ultimately, contribute to improve the coverage and metadata completeness of the considered data sources by enriching both local and global repositories. Such an approach emerges in a context where some open global repositories make their code for selection and indexation freely available, which allows not only for more transparency than commercial alternatives, but also allows to directly contribute to the improvement by identifying gaps. In contrast to strictly descriptive approaches, this curative approach has the side effect that it may render the results of an analysis obsolete shortly after being performed, but with the benefit that it can improve the considered data sources.

Conclusion and Future Directions

To identify what is missing when relying solely on a local or global repository, we propose a DOI Screening System that provides rapid overlap analysis and suggests "curative" actions to improve or interpret mismatches. The tool requires minimal input: an institutional identifier, a list of DOIs, and a time range. Since the system's code is publicly available, we expect community-driven enhancements to refine and expand the predefined rules, thereby increasing the portion of matched publications. Our immediate plans include extending the DOI Screening System, that currently only handles OpenAlex as a global repository, by incorporating OpenAIRE. We will also deepen the existing case studies, and explore how future iterations of the system could handle other units of analysis, such as a researcher's ORCID or a journal's ISSN. By embedding rules that specifically address gaps in the overlap analysis, the DOI Screening System serves as a catalyst for enhancing both the coverage and quality of local and global repositories, ultimately fostering more effective dissemination of scientific publications.

Acknowledgments

Contributors from the library of ETH Zurich worked on this paper as part of the TOBI project (Towards Open Bibliometric Indicators), co-funded by swissuniversities and the ETH Library.

References

Alperin, J. P., Portenoy, J., Demes, K., Larivière, V. & Haustein, S. (2024). An analysis of the suitability of OpenAlex for bibliometric analyses. *arXiv*. https://doi.org/10.48550/arXiv.2404.17663.

Andreose, E., Di Marzo, S., Heibi, I., Peroni, S. & Zilli, L. (2025). Analysing the coverage of the University of Bologna's publication metadata in an existing source of open research information. *arXiv*. https://doi.org/10.48550/arXiv.2501.05821.

Bologna, F., Di Iorio, A., Peroni, S. & Poggi, F. (2022). Open bibliographic data and the Italian National Scientific Qualification: measuring coverage of academic fields. *Quantitative Science Studies*, 3 (3), 512–528. https://doi.org/10.1162/qss_a_00203.

Delgado-Quirós, L., Aguillo, I. F., Martín-Martín, A., López-Cózar, E. D., Orduña-Malea, E., & Ortega, J. L. (2024). Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases. *Journal of the Association for Information Science and Technology*, 75(1), 43–58. https://doi.org/10.1002/asi.24839.

Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A. & de Moya-Anegón, F. (2021). Comparative analysis of the bibliographic data sources Dimensions and Scopus: An approach at the country and institutional levels. *Frontiers in Research Metrics and Analytics*, 5. https://doi.org/10.3389/frma.2020.593494.

Hug, S.E., Ochsner, M. & Brändle, M.P (2017). Citation analysis with Microsoft Academic. *Scientometrics*, 111, 371–378. https://doi.org/10.1007/s11192-017-2247-8.

Hug, S.E. & Brändle, M.P. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, 113, 1551–1571. https://doi.org/10.1007/s11192-017-2535-3.

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: a multidisciplinary comparison of coverage via citations. *Scientometrics*, 126, 871–906. https://doi.org/10.1007/s11192-020-03690-4.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: a fully-open index of scholarly works, authors, venues, institutions, and concepts. ArXiv. https://arxiv.org/abs/2205.01833.

Vieira, G.A. & Leta, J. (2024). *Biblioverlap*: an R package for document matching across bibliographic datasets. *Scientometrics*, 129, 4513–4527. https://doi.org/10.1007/s11192-024-05065-5.

FULL PAPER

https://doi.org/10.51408/issi2025_022

"From Essential to Obsolete? The Evolution of Personal Communications in Academic Research and Citation Practices"

Juan Gorraiz

juan.gorraiz@univie.ac.at
University of Vienna, Vienna University Library, Boltzmanngasse 5, A-1090 Vienna (Austria)

Abstract

This study examines the evolving role of personal communications in academic research, tracing their historical significance and transformation in citation practices. Traditionally encompassing verbal exchanges, letters, and private correspondence, personal communications have long served as valuable but non-retrievable sources of knowledge. Using Scopus bibliometric data (1971–2024), this study investigates citation trends, disciplinary differences, and the growing impact of digitalization and artificial intelligence (AI) on informal scholarly exchanges. Findings indicate a decline in personal communication citations since the 2000s, likely due to the rise of formalized digital documentation, preprints, and AI-assisted research tools.

However, certain disciplines—such as Social Sciences and Computer Science—continue to rely heavily on personal communications, underscoring their ongoing relevance. The study also highlights a significant gap in citation standards, particularly in cases such as peer review reports, where proper attribution remains undefined. Furthermore, the potential classification of AI-generated insights as a form of personal communication raises new questions about citation ethics and research transparency. This pilot study contributes to bibliometric research by mapping the evolution of personal communications and advocating for standardized citation practices that reflect contemporary academic exchanges.

Introduction

The practice of citing personal communications holds a unique place in academic and scientific discourse. Historically, such communications have encompassed direct verbal exchanges, written correspondence (e.g., letters), and informal discussions, often occurring spontaneously at conferences or meetings. These exchanges, though not formally published, have played a crucial role in shaping scientific knowledge.

In the early modern period, correspondence between scholars served as a precursor to modern peer review, allowing researchers to share findings and experimental methods with colleagues or members of scientific societies (Gross et al., 2002). Even with the establishment of journal-based scholarly communication systems in the 17th century (Manten, 1980), informal exchanges remained vital to intellectual progress. Over time, these communications evolved, taking various forms, including direct verbal exchanges (such as personal interviews and discussions at academic events) and written correspondence (such as letters and emails). Letters, in particular, have been invaluable for historical research, while emails—though private—are

frequently cited as they serve as direct records of academic exchange (Cronin & Franks, 2006).

As scholarly communication systems developed, citation guidelines sought to standardize the inclusion of informal sources. Style guides such as the Chicago Manual of Style (first published in 1906) and the American Psychological Association (APA) Style Manual (introduced in 1929) began addressing how personal communications should be integrated into academic work. By the mid-20th century, APA explicitly instructed that personal communications—including unpublished letters, verbal exchanges, and private emails—should be cited only within the text and omitted from reference lists. This practice, formalized in the APA Style Manual's first edition (1952), remains in place today. Similarly, contemporary publishers, including Elsevier, specify that "unpublished results" and "personal communications" must adhere to standard reference styles, typically replacing publication dates with these terms (Day, Gastel, & Buchanan, 2012). This study focuses on the case of "personal communication," distinguishing it from "unpublished and negative results."

The nature of personal communication in academia has evolved significantly due to two major forces: the widespread adoption of the internet (Longo et al., 2009) and advancements in artificial intelligence (Dwivedi et al., 2021). Since the 1970s, the expansion of digital technologies and the growing emphasis on academic collaboration have led to an increase in multi-authored works (Brand et al., 2015). This, in turn, has broadened the concept of personal communication beyond one-on-one interactions to include diverse forms of exchange, such as emails, social media discussions, and online forums (Kousha, Thelwall, & Abdoli, 2012). These new communication channels blur the lines between formal publications and informal knowledge-sharing, raising questions about how such exchanges should be cited and acknowledged in scholarly work.

Two recent developments highlight the need to revisit citation practices for personal communications:

1. Plagiarism in Peer Review: A recent case of plagiarism during the review process of a scholarly manuscript exposed gaps in current citation standards. The plagiarized material, derived from a reviewer's comments, did not fit neatly within existing citation guidelines. While such content might be classified as personal communication, the absence of explicit standards creates ambiguity—especially in peer-review contexts (Ross-Hellauer, Deppe, & Schmidt, 2017).

2. The Role of AI in Academic Communication: The expansion of AI-generated content introduces new challenges in citation norms. A recent study (Gorraiz, 2025) examined the role of AI tools (e.g., ChatGPT) in academic research, particularly investigating whether they are recognized as authors or co-authors and how their contributions are cited across disciplines. Given that AI-generated outputs often function as sources of information—providing insights that are not directly retrievable—there is increasing interest in contextualizing AI citations within the broader framework of personal communications (Haustein et al., 2023).

Traditionally, personal communications have facilitated scholarly exchange by allowing researchers to share insights, theories, and unpublished data through

informal channels such as correspondence, interviews, and discussions. AI-generated outputs, which provide non-retrievable but influential knowledge, could be seen as analogous to these traditional forms of exchange. However, citation practices for AI remain inconsistent and largely unstandardized. This raises important questions: Should AI-generated insights be classified under personal communications? If expert discussions and peer exchanges qualify as valid informal sources, could AI outputs be acknowledged in the same way? As academic communication becomes increasingly structured, will AI tools replace traditional human-mediated informal exchanges, reshaping the landscape of personal communications?

Objective of the Study

While research has extensively examined citation patterns, co-authorship dynamics, and academic communication, the practice of citing *personal communications* remains underexplored in bibliometric studies. This paper aims to fill this gap by investigating the evolution of personal communication citations in scientific literature and examining their representation in bibliometric databases.

Research Questions

1. How are personal communications cited in scientific literature? 'This question explores if personal communications are cited in the reference lists or not,

2. Do bibliometric databases track personal communications in citations? Major databases like Web of Science and Scopus are essential tools for tracking citations, but do they accurately capture personal communications? Given the lack of standardization in citing these sources, this study investigates whether and how they can be identified and analyzed.

3. How have citations of personal communications evolved over the past decades? This question examines historical trends, focusing on how digitalization and the rise of online platforms (emails, blogs, social media) have impacted their citation. Has the increased accessibility of digital communications led to greater or lesser reliance on personal communications, and how have citation practices adapted?

4. In which academic fields are personal communications most commonly cited? This question aims to identify the disciplines where personal communications are frequently cited. Are they more prevalent in social sciences, humanities, or STEM fields?

By mapping the historical development of personal communication citations, this first pilot study aims to establish a foundation for understanding their current role and the challenges posed by emerging technologies. Examining the intersection of AI and personal communications will provide valuable insights into how informal knowledge-sharing is evolving in response to technological advancements and shifting academic norms. Future research will expand on this initial analysis, exploring disciplinary differences and their implications for academic integrity in the digital age.

Methodology

Initially, the analysis was planned to include the two largest and oldest scientometric databases: Web of Science Core Collection (WoS CC) and Scopus. However, serious difficulties were encountered while downloading and cleaning the data from WoS CC. These challenges resulted in outputs of questionable validity, which prompted us to restrict this preliminary study to the Scopus database. The decision to focus on Scopus was based on its clearer and more interpretable methodological framework. The choice to defer the integration of WoS CC data to a future study was made to ensure the reliability of the findings. A subsequent investigation will focus on assessing the capacity and suitability of WoS CC for measuring citations to personal communications, an issue that remains open and warrants further exploration.

In Scopus, the search string REF("pers* comm*") was used in the Advanced Search.This search yielded 232,429 documents that cited one of these terms in their references (as of 1.11.2024). As two indexed and cited journals, IEEE Personal Communications[1] and Wireless Personal Communications[2], were found under the results of this search in Scopus, they were excluded from this analysis. Thus, the refined search string to identify citing documents was: `((REF("person* commun*")) AND NOT (REF("wire* person* commun*")) AND NOT (REF("IEEE* person* commun*"))`. This search returned 95,580 citing documents. Due to download limits (max. 20,000 documents), the data was downloaded in batches organized by publication year.

To retrieve cited documents, the following steps were taken: Within the above described search, the "Secondary documents"[3] tab was activated to identify documents referenced in Scopus articles but not directly available in the Scopus database. 76,538 documents were obtained (as of 1.11.2024). The search was then refined to include only documents where the source contained any form of "person* commun*" (i.e. again, documents citing the two journals IEEE Personal Communications and Wireless Communications were excluded from further analysis). The remaining 65,544 documents (approximately 85% of the initial amount) were then analysed. Cited personal communications were clustered according to their citation form. Most common citation forms were identified and depicted.

To clarify the terminology used in this study, we distinguish between **cited personal communication** and **citing personal communication** as follows:

---

[1] IEEE Personal Communications ceased publication in 2001. The current retitled publication is **IEEE Wireless Communications**.

[2] **Wireless Personal Communications** is an archival, peer reviewed, scientific and technical journal addressing mobile communications and computing. It investigates theoretical, engineering, and experimental aspects of radio communications, voice, data, images, and multimedia. The journal features five principal types of papers: full technical papers, short papers, technical aspects of policy and standardization, letters offering new research thoughts and experimental ideas, and invited papers on important and emerging topics authored by renowned experts.

[3] According to Scopus, a **secondary document** is "a document that has been extracted from a Scopus document reference list but is not available directly in the Scopus database since it is not indexed by Scopus." For these secondary documents, limited functionality is available.

- **Cited personal communication** refers to any reference explicitly labeled as "personal communication" in a Scopus-indexed journal. Since personal communications cannot constitute a source document (i.e., they are not formally published works), they only appear under the category of *secondary documents* within the Scopus database.
- **Citing personal communication** refers to any article indexed in Scopus that, in its year of publication, has cited at least one "personal communication." It is important to note that personal communications can only be cited in the same year or in previous years relative to the publication date of the citing article, as they lack a formal publication timeline.

The evolution of the number of cited personal communications and the number of citing articles was then retrieved. To assess whether the evolution of citations of personal communications is solely influenced by the increasing number of publications indexed in this source, the annual number of indexed publications in Scopus was retrieved using the Advanced Search feature and the command 'PY after 1970.' By dividing the annual number of cited and citing personal communications by the number of publications indexed each year, we calculated the "Normalized Citation Rate (NCR)" o "Normalized Citation Frequency (NCF)" of personal communications in the data source Scopus. To facilitate interpretation, this value was multiplied by 10,000. The resulting metric represents the normalized citation rate per 10,000 publications indexed annually for cited PCs or documents citing PCs.
To address Research Question 3, which investigates how personal communications are cited in academic literature and the contexts in which they appear, a series of systematic searches were conducted in the Scopus database. These searches aimed to identify instances where the phrase "personal communication" (or its variations) was used in conjunction with specific terms that represent various forms of communication.
The queries employed Boolean logic with proximity operators to ensure that relevant terms appeared within close context (7 words apart) of the key phrase. The following search strings were used:

- REF("person* commun*" W/7 "oral*") – to identify citations referencing oral communications.
- REF("person* commun*" W/7 email*) – for emails.
- REF("person* commun*" W/7 letter*) – for written letters.
- REF("person* commun*" W/7 interview*) – for interviews.
- REF("person* commun*" W/7 meeting*) – for meetings.
- REF("person* commun*" W/7 conference*) – for conferences.
- REF("person* commun*" W/7 memo*) – for memos.
- REF("person* commun*" W/7 blog*) – for blogs.
- REF("person* commun*" W/7 openai*) – for OpenAI tools.
- REF("person* commun*" W/7 chatgpt*) – for ChatGPT references.
- REF("person* commun*" W/7 review*) – for reviews.

These searches performed on November 20224 allowed us to explore how personal communications are contextualized in academic citations, particularly in relation to

oral and written forms of communication, emerging AI tools like ChatGPT, and specific collaborative settings such as meetings or conferences. The findings from these targeted searches were analyzed to determine the prevalence of personal communications in different contexts and their alignment with citation practices in the scholarly literature.

Finally, citing articles were grouped into subject areas, and we ranked subject areas according to the ratio of citing personal communications within each field and thus identified disciplines where personal communications seem to play a prominent role. These results were compared with the percentages each area represented in the database during the analyzed period (after 1970) to determine whether the proportions merely reflect the coverage of each discipline within the database.

Results

Table 1 below lists the most common citation forms for "personal communications" in Scopus. The most frequently used form is "Personal communication," with over 29,000 citations, followed by "Personal Communications" with 1,167 citations. Variations in formatting, such as capitalization, punctuation, and inclusion of phrases like "to the author" or "via email," create a wide range of forms. This diversity in citation styles reflects inconsistency in how personal communications are referenced across different documents in Scopus.

Table 1. Most common citation forms for "Personal Communications" in Scopus.

Cited form in Scopus	# citations	% of 65544
Personal communication	29261	44.64%
Personal Communications	1167	1.78%
Personal communication.	813	1.24%
Personal communication with the author	386	0.59%
Personal communication with author	317	0.48%
Personal Commun	154	0.23%
Personnal Communication	134	0.20%
Personal Communication to the Author	104	0.16%
Personnel communication	95	0.14%
Personal communication to author	77	0.12%
Personal Communication With the Authors	74	0.11%
Personal Communication Via Email	59	0.09%
Personal Commun.	59	0.09%
Personal Communication,	53	0.08%
Personal communications.	52	0.08%
Personal communication with authors	41	0.06%
Personal Communication Via E-mail	35	0.05%
Personal communication by email	35	0.05%
(Personal Communication)	33	0.05%
Personal Communication to the Authors	32	0.05%

The results of the evolution of the cited and citing personal communications in Scopus are shown in **Figure 1**. The number of "personal communications" citations in Scopus reveals a notable pattern: Those citations began to gain momentum in the early 1970s, corresponding with a period when Scopus's coverage became more comprehensive. This growing trend in absolute numbers of "personal communications" citations continued, reaching a peak around **2012** with approximately 2,215 cited references and 5,000 citing documents. Post-2018, a decline in citations is apparent, suggesting a reduced emphasis on "personal communications" as a source in scientific literature.

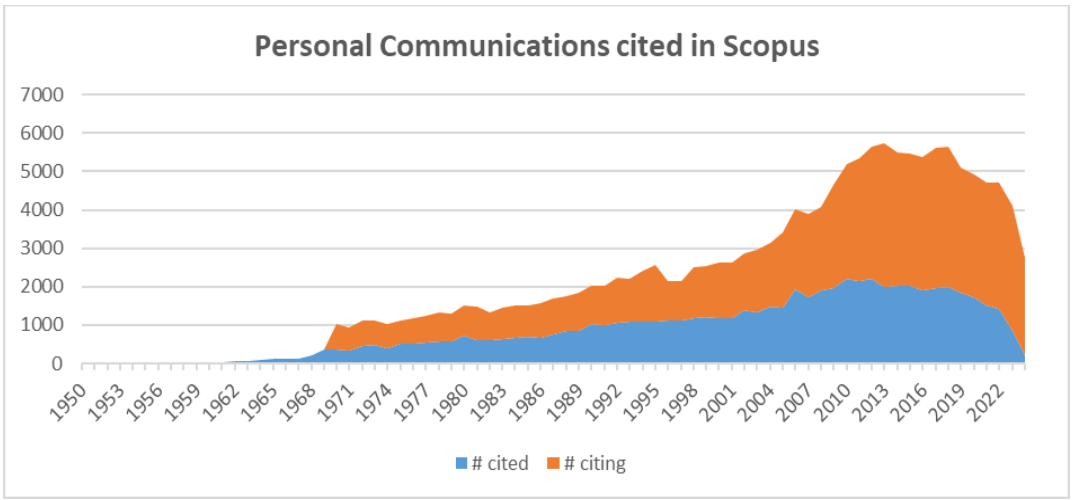


Figure 1. Trends in Citations/Cited of "Personal Communications" in Scopus.

Figure 2 shows the results for the Annual Normalized Citation Rate (NCR) of Personal Communications in Scopus (*cited* in red; *citing* in blue). This normalization eliminates potential effects caused by annual variations in the number of publications indexed in the Scopus database, ensuring a more accurate comparison over time.

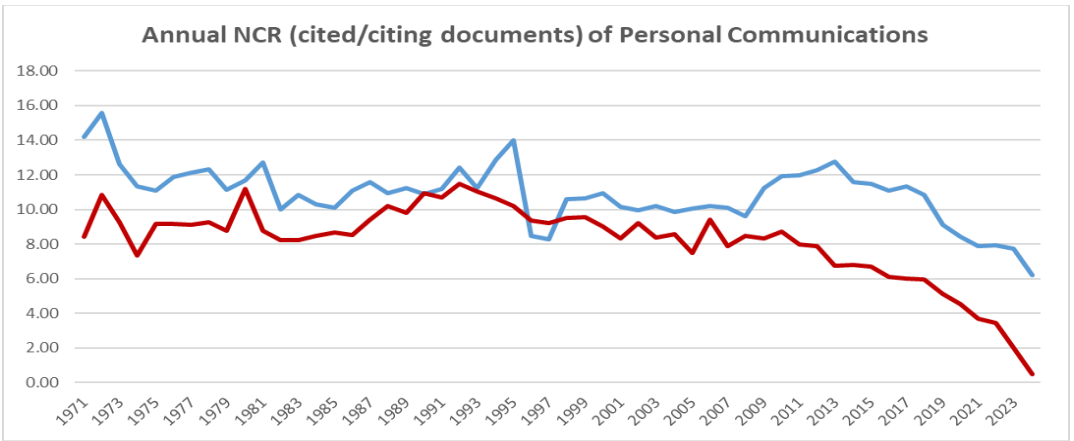


Figure 2. Annual Normalized Citation Rate (NCR) of Personal Communications in Scopus (cited in red; citing in blue).

From this chart, the following insights can be drawn:

1. Overall Decline in Citations of Personal Communications:
   - The graph shows a general decline in the annual NCR (Normalized Citation Rate) of personal communications over the decades.
   - While both the cited and citing trends started relatively high in the 1970s, they have consistently decreased, with a sharper decline after the late 1990s.
2. Sharp Drop in the Late 1990s and Early 2000s:
   - The late 1990s and early 2000s saw a significant decline in the use of personal communications as sources in citations. This period coincides with the rise of digital communication platforms, particularly the increased adoption of email and the early stages of the internet becoming widely available.
3. Impact of Social Media and Digitalization:
   - The continued decline through the 2010s aligns with the rise of social media platforms, blogging, and other online platforms that may have replaced informal personal communications as a source for scholarly interaction. Digital platforms offer more public, archivable, and citable forms of communication, potentially reducing reliance on private and informal exchanges.
4. Steepest Decline in the Past Decade (2010-2020):
   - The steep decrease in citations during this period may reflect a paradigm shift in scholarly communication. Researchers might prefer more formal and traceable sources, such as public online discussions, preprints, or data repositories, over informal personal communications.

Table 2. Results of citation of personal communications in different contexts.

Search query	# items	# secondary documents	cited by
REF("person* commun* W/7 "oral*)	39	41	26
REF("person* commun*" W/7 email*)	775	576	497
REF ("person* commun*" W/7 letter*)	356	141	118
REF ("person* commun*" W/7 interview*)	344	567	306
REF ("person* commun*" W/7 meeting*)	273	189	194
REF ("person* commun*" W/7 conference*)	9972	298	
REF("person* commun*" W/7 memo*)	319	32	44
REF ("person* commun*" W/7 blog*)	5	3	3
REF ("person* commun*" W/7 openai*)	4	1	1
REF ("person* commun*" W/7 chatgpt*)	5	4	5
REF("person* commun*" W/7 review*)	1804	304	151

These results illustrate the prevalence of personal communications in different contexts:

- The most frequent context for citing personal communications was found in "review" documents, with 1,804 items, followed by "conference" documents (9,972 items) and "email" communications (775 items). This indicates that personal communications are most often referenced in reviews and conferences, suggesting these contexts emphasize informal or non-formalized exchanges.
- AI-Related Citations:
  Emerging technologies like ChatGPT and OpenAI tools showed minimal representation, with only 4 items each. This suggests that, at present, AI tools have a limited role in personal communications as cited in academic work, which may change as these technologies become more integrated into scholarly activities.
- Secondary Documents and Citations:
  Secondary documents and citation counts were relatively consistent with the prevalence of the primary items. For instance, "email" communications were referenced in 576 secondary documents and cited 497 times, while "letter" communications were associated with 141 secondary documents and cited 118 times. The high citation count for email communications highlights its growing importance as a medium of exchange in academia.
- Interpersonal Communication Forms:
  Other forms of interpersonal communication, such as meetings (273 items), interviews (344 items), and memos (319 items), demonstrated moderate representation. However, their relatively low secondary document and citation counts suggest that these contexts are not as widely disseminated or influential as conference or review materials.
- Comparison Across Contexts:
  Interestingly, oral communication was cited 39 times, with relatively low representation in secondary documents (41) and citations (26). This may indicate that oral communications are harder to formalize or verify in academic publications compared to written or electronic exchanges. Similarly, blogs (9 items) showed limited relevance in academic references.

Finally, **Table 2** presents an analysis of the subject areas in which "personal communications" are most frequently cited. This analysis was conducted on the 95,580 citing documents between 1971 and 2024, revealing the disciplines where personal communications play a prominent role in reference practices.

Table 2. Top Scopus Subject Areas Citing Personal Communications (1971.2024).

SUBJECT AREA	# Publications citing Personal Communications	% citing publications	Subject Area Percentage in Scopus	Normalized % publications citing
Engineering	23330	14.27%	11.96%	1.19
Computer Science	18955	11.60%	6.24%	**1.86**
Social Sciences	17925	10.97%	5.34%	**2.05**
Medicine	15835	9.69%	17.31%	0.56
Mathematics	9351	5.72%	3.84%	1.49
Environmental Science	9067	5.55%	3.30%	**1.68**
Chemistry	7609	4.66%	4.89%	0.95
Arts and Humanities	6715	4.11%	2.75%	1.49
Physics and Astronomy	6564	4.02%	6.88%	0.58
Agricultural and Biological Sciences	5976	3.66%	4.00%	0.91
Materials Science	5167	3.16%	5.80%	0.55
Biochemistry, Genetics and Molecular Biology	5112	3.13%	6.80%	0.46
Earth and Planetary Sciences	4750	2.91%	2.75%	1.06
Energy	4197	2.57%	1.99%	1.29
Psychology	4052	2.48%	1.36%	**1.82**
Business, Management and Accounting	3617	2.21%	1.59%	1.39
Chemical Engineering	3116	1.91%	2.61%	0.73
Economics, Econometrics and Finance	2398	1.47%	1.06%	1.38
Others	9714	5.94%		

The table presents the distribution of publications citing personal communications across different subject areas in the Scopus database (1971–2024). The Normalized % Publications Citing column adjusts for the representation of each subject area within the database, providing a more accurate comparison.
Key findings include:

- Social Sciences (2.05), Computer Science (1.86), and Psychology (1.82) exhibit the highest normalized citation rates for personal communications. These fields rely significantly on informal and direct exchanges, possibly due to their emphasis on qualitative insights, theoretical discussions, and evolving methodologies.
- Engineering (1.19), Mathematics (1.49), and Environmental Science (1.68) also show above-average reliance on personal communications, indicating that these disciplines often engage in direct knowledge-sharing beyond formally published literature.
- In contrast, Medicine (0.56), Biochemistry, Genetics and Molecular Biology (0.46), and Materials Science (0.55) have notably low normalized citation rates. These fields generally depend more on formal, peer-reviewed sources, where reproducibility and documentation are crucial.

- Physics and Astronomy (0.58) and Chemistry (0.95) also exhibit lower-than-average reliance, likely due to the structured and empirical nature of their research.

Discussion and conclusions

Despite the fact that major style guides continue to stipulate that personal communications should be cited only in the text and not included in the reference list, our results obtained from Scopus show a significant use of personal communications as references, with notable differences in citation formats and volume. The presence of abbreviations and minor format variations highlights the lack of standardized citation practices for personal communications. These inconsistencies can distort bibliometric analyses, leading to under- or over-representation of certain citation forms.

Given this variability, we propose the adoption of a standardized citation format for personal communications: **Cited Person, Year, Personal Communication: Type (e.g., oral, letter, memorandum, interview, email, social media, etc.).**

Such a format would enhance data consistency and comparability in bibliometric studies while preserving transparency in academic referencing.

The findings of this study are a first attempt to highlight the evolving role of personal communications in scholarly research and how their citation patterns vary across disciplines. The results reveal a significant shift in the use of personal communications as citations over time, reflecting broader transformations in academic communication, technological advancements, and changing publication practices. However, the observed decline in such citations is not due to a stricter enforcement of these long-standing style guidelines—there is no evidence to suggest this—but rather to the transformative impact of technological advancements in scholarly communication.

The study results hints at a progressive decline in personal communication citations, particularly since the late 1990s, corresponding with technological advancements and digitalization in scholarly communication. The sharp drop observed in the early 2000s coincides with the widespread adoption of email, digital archives, and open-access repositories, which have provided researchers with more formalized, traceable, and archivable alternatives to personal communications. The further decline in the 2010s and 2020s aligns with the emergence of preprint servers, academic social networks, and AI-generated research tools, which enable rapid knowledge dissemination without relying on direct personal exchanges (Koutras, 2021).. These findings might indicate that as scholarly communication becomes more structured, informal references are becoming less relevant in academic citations. However, informal exchanges themselves remain central to knowledge production, even if they are less frequently acknowledged in citation records.

The subject-area analysis shows substantial variation in the reliance on personal communications across disciplines:

- Social Sciences: The high citation rates likely stem from the importance of qualitative insights, interviews, and theoretical discussions, which often rely on informal exchanges rather than strictly published sources.
- Computer Science: The strong reliance on personal communications may reflect the field's rapidly evolving nature, where many breakthroughs first circulate through direct peer discussions before formal publication.
- Medicine and Biochemistry: These fields follow highly structured research methodologies, where reproducibility and verification are critical, reducing the necessity for citing informal communications.
- Physics, Chemistry, and Engineering: These disciplines show moderate reliance on personal communications, potentially due to collaborative work environments where technical discussions and experimental insights are shared informally before publication.

These differences highlight how personal communications are perceived and utilized differently depending on the academic field.

The Changing Role of Informal Communication in Academia

Although citations of personal communications have declined, informal academic exchanges remain central to research collaboration. The transition toward digital platforms, AI-driven tools, and collaborative research networks is reshaping how scholars share knowledge.

Interestingly, the low representation of AI-related citations (e.g., ChatGPT and OpenAI, with only four citations each) suggests that AI-generated insights are not yet widely recognized as a valid form of personal communication in academia. However, this trend may be shifting. As highlighted in recent bibliometric analyses, AI tools are increasingly being cited as sources or acknowledged in research papers, reflecting their growing role in scientific discourse, despite the lack of formal authorship recognition (Gorraiz, 2025).

These preliminary results answer the question "Should AI-generated insights be classified under the same category as personal communications?" with a clear no. AI-generated insights are not cited under the category of personal communication but rather follow their own distinct citation dynamics.

The findings of Gorraiz (2025) indicate that while AI contributions are still in an early adoption phase, their presence is expanding, particularly through acknowledgments and citations as computational tools. Ethical concerns and academic publishing guidelines (e.g., COPE) currently prevent AI from being credited as an author, reinforcing the notion that AI is primarily viewed as an assistive tool rather than an intellectual contributor. However, as AI tools become more embedded in scholarly workflows, their influence on informal academic exchanges and citation practices is expected to grow substantially, potentially reshaping how researchers engage with personal communications in the future.

The continued prevalence of emails, letters, and direct correspondences in reviews and conference proceedings suggests that despite the decline in citation frequency, personal communications still play a significant role in academic knowledge-sharing. Review articles and conference papers frequently cite unpublished

conversations, expert opinions, and preliminary results, reinforcing the idea that informal exchanges are still valuable, even if they are less frequently acknowledged in citation records.

Addressing Citation Challenges in Peer Review and AI-Generated Content

One possible solution to citation inconsistencies is to categorize peer review comments under personal communications, acknowledging the reviewer as the source. This approach would align with ethical academic standards and ensure proper recognition of intellectual contributions. The absence of clear citation guidelines for peer review content has thus emerged as a key motivation for this study, highlighting the urgent need for publishers and institutions to develop standardized recommendations that promote transparency and respect within the peer review process (Tennant et al., 2019).

Finally, this study suggests that AI-generated insights are not yet widely cited as personal communications but may soon become more prominent. Future research should further explore the evolving role of AI-generated knowledge in academic citations and investigate whether AI-driven tools will transform informal scholarly exchanges.

Final Thoughts and Future Considerations

By mapping the historical trajectory of personal communications as citations, this study provides a foundation for understanding their current role and the challenges posed by emerging technologies. As digital communication continues to evolve, the boundaries between formal publications and informal scholarly exchanges will likely continue to shift, shaping the future of academic discourse.

This study represents the first in-depth attempt to analyze the evolution of personal communications in scientific discourse and is part of an ongoing research project at the University of Vienna. As such, the findings should be viewed as preliminary insights, with further analyses planned to assess the suitability of data sources and provide a deeper contextual interpretation of the results. In parallel, we are also investigating whether personal communications are mentioned in acknowledgments, and these results will be presented at the upcoming conference.

These findings also highlight the importance of rigorous data handling in bibliometric research, particularly when analyzing citation forms with high variability. Researchers utilizing bibliometric databases should be aware of inconsistencies and potential indexing errors to ensure accurate representation and interpretation of citation trends in personal communications.

Finally, this study underscores how disciplinary differences, technological advancements, and the open-access movement influence how personal communications are incorporated into academic research. As digital communication continues to evolve, the boundaries between formal publications and informal scholarly exchanges will likely continue to shift, shaping the future of academic discourse.

Limitations

Despite these efforts to clean and refine the data, limitations inherent to the databases and their search functionalities may still have influenced the findings.

Another limitation of this study, which is common in scientometric and sociological research, is the lack of a strong cause-and-effect relationship. One of the primary reasons for this is the inherent inability to eliminate all other potential causal factors from the analysis. Consequently, particularly with regard to Research Question 3, our findings can only point to **signs** that require further observation and investigation to be fully confirmed. For instance, the idea that personal communications have already been replaced by the internet, which has established its own specific channels of communication—from emails to blogs, among others—and might be further overshadowed by the rise of AI tools, can only be suggested as a potential trend. Similarly, the notion that the esteemed and trusted colleague will not eventually be replaced by an intelligent tool—one built on the knowledge and experience of countless professionals—appears to be more a matter of time than an impossibility. However, these observations remain speculative and require longitudinal studies to validate such hypotheses.

Acknowledgments

I would like to express my sincere gratitude to my colleague at the University of Vienna, Iwona Dullinger, for her invaluable assistance in formulating the introduction and conducting the reference search. Her insightful ideas, critical feedback, and thoughtful suggestions have undoubtedly contributed significantly to improving this contribution presented for the conference.

References

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. Learned Publishing, 28(2).

Cronin, B., & Franks, S. (2006). Trading cultures: Resource mobilization and service rendering in the digital library economy. Library Trends, 54(4), 748-765.

Day, R. A., Gastel, B., & Buchanan, R. (2012). How to write and publish a scientific paper. Cambridge University Press.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International journal of information management*, *57*, 101994.

Gorraiz, J. (2025). Acknowledging the new Invisible Colleague: Addressing the Recognition of Open AI Contributions in in Scientific Publishing. *Journal of Informetrics,* in press

Gross, A. G., Harmon, J. E., & Reidy, M. S. (2002). Communicating science: The scientific article from the 17th century to the present. Oxford University Press, USA.

Haustein, S., Larivière, V., Thelwall, M., Amyot, D., & Peters, I. (2023). Academics' use of social media: Implications for scholarly communication. PLoS ONE, 18(3), e0263257.

Kousha, K., Thelwall, M., & Abdoli, M. (2012). The role of online scholarly resources in the university research process: An analysis of researchers' e-resource access and use. Aslib Proceedings, 64(2), 162-177.

Koutras, N. (2021). The rise of preprints: Implications for the future of peer-reviewed scientific publishing. Scientometrics, 126(8), 6291-6310.

Longo, M., & Magnolo, S. (2009). The Author and Authorship in the Internet Society: New Perspectives for Scientific Communication. *Current Sociology*, *57*(6), 829-850. https://doi.org/10.1177/0011392109342221

Manten, A. (1980). Publication of scientific information is not identical with communication. *Scientometrics*, 2(4), 303-308.

Ross-Hellauer, T., Deppe, A., & Schmidt, B. (2017). Open peer review: Promoting transparency in open science. F1000Research, 6, 588.

Tennant, J. P., Dugan, J. M., Graziotin, D., Jacques, D. C., Waldner, F., & Mietchen, D. (2019). A multi-disciplinary perspective on emergent and future innovations in peer review. F1000Research, 6, 1151.

https://doi.org/10.51408/issi2025_041

A Comparative Study on Text Multi-Features Mining for Patent Text Clustering: The Case of Graphene Sensing Technology

Xian Zhang[1], Jiahui Li[2], Shuying Li[3], Haiyun Xu[4]

[1]*zhangx@clas.ac.cn*, [2]*lijiahui222@mails.ucas.ac.cn*, [3]*lisy@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

[4]*xuhaiyunnemo@gmail.com*
Business School, Shandong University of Technology, Zibo (China)

Abstract

The development of text feature extraction and measurement methods has given rise to a diversification of perspectives on text mining. However, few studies have explored the similarity, complementarity, and effectiveness of different text features. The selection of different feature combinations lacks a supporting basis. This study selected four types of text feature words from patent texts, namely, text domain feature keywords by Comprehensively Measure Feature Selection algorithm (CMFS), technical interdisciplinary keywords by the term Interdisciplinary index (TI), technical breakthrough keywords by the Kleinberg burst detection algorithm (KB) and high-frequency words (HF). A set of measurement indicators and implementation methods based on the Jaccard distance index, information entropy, and mutual information theory was designed, to determine the similarities, differences, synergies, and complementarities of the four types of text feature words. Based on comparative analysis, a comprehensive measurement index was designed, as well as feature combinations were selected. To illustrate this approach, we selected patent documents in the domain of graphene sensing and evaluated various feature combinations with different word embedding and clustering algorithms. The results show that multivariate features enhance the effectiveness of single high-frequency features in text mining tasks. There is a wide range of applicability for CMFS+KB feature combination, with the clustering effect being optimal when used with FsatText+K-means. For the specific case of HDBSCAN+FastText, the HF+CMFS+KB feature combination demonstrates superior performance. This study corroborates the information representation significance and complementarity of four types of keywords in information representation, thereby substantiating the extraction and analysis of text multi-features. Finally, we also point out the limitations of measurement methods and feature types in the research and prospects for future research.

Introduction

Text feature words refer to keywords that can represent the main theme, meaning, content and other features of the text. They are widely used in fields such as information retrieval and text classification (Chi et al., 2019). Text feature words are usually extracted directly from the text. Word frequency represents the most widely applied fundamental method for extracting text feature words. This method reveals the text topic by analyzing and describing lexical rules (Feng Guohe & Kong Yongxin, 2020; Salton, Allan, & Singhal, 1996). For example, high-frequency words often dominate in topic classification and identification (Li, Zhang, Li, Ouyang, & Cai, 2018).

However, in the field of patent text mining, topic models often tend to favor high-frequency words and have limitations in implicit semantic expression (Yu Yan & Zhao Naixuan, 2018). Cassandra L. Jacobs et al. (Jacobs, Dell, Benjamin, & Bannard, 2016) also proposed that high-frequency words are more easily recognized in cognitive processes, while low-frequency words exhibit enhanced recognizability and potentially contain more significant information. Therefore, beyond the comprehensive mining of high-frequency features, there has been a surge of interest within the academic community in the mining of selecting low-frequency words as a complement to high frequency words. In addition, multi-feature extraction has been found to be more conducive to the accuracy of machine understanding for text mining (Cheng Yong, Xu Dekuan, & Lv Xueqiang, 2019). The perspective of extracting text feature words is constantly enriched, such as revealing important features of the field, technical interdisciplinary features, and technical breakthrough features, which have been widely used in research.

On the one hand, there are few studies on how the words extracted from different feature relationships represent the text topics; the similarities and differences between these representations; the potential supplementary role of these representations for high-frequency words; and how to quantitatively measure the differences in their information meaning. On the other hand, few researchers have explored the practical effects of different combinations of multiple features in text mining tasks.

This study aims to address this gap by conducting quantitative measurement and comparative research on the text topic representation effects of different types of text feature words. The objective is to provide scientific and quantitative reference for text topic feature mining. To illustrate this, we selected four types of text feature words from patent texts, namely, text domain feature keywords (CMFS), technical interdisciplinary keywords (TI), technical breakthrough keywords (KB) and high-frequency words (HF). We then designed a set of comprehensive measurement indicators for feature combinations and implementation methods based on similarity, information entropy, and mutual information theory. A comparative study was conducted on the similarity, difference, complementarity, and synergy of the text representations of four types of text feature words. Different feature combinations were applied to three word embedding models and three clustering algorithms to explore the application effect of multi-feature combination in text clustering.

The rest of this study is as follows: Section 1, Introduction; In Section 2, we reviewed the application of word embedding and clustering algorithms, as well as the extraction and selection methods of different text feature words. Based on this design, the research framework is obtained; In Section 3, we introduced the data source and vocabulary extraction. We also propose comparative analysis methods, comprehensive indicator design, word embedding model and clustering algorithm; In Section 4, we present the empirical results and analysis; In Section 5, we summarize the characteristics and usage scenarios of multi-feature combinations based on the results; Finally, we summarize the theoretical and practical significance of this study, as well as its limitations and future research directions.

Literature review

Word embedding and text clustering

In text mining tasks, vocabulary is the core unit and the basic representation form of knowledge content in the semantic field. With the maturity of word representation technology in natural language processing, existing research has mostly used word embedding models to generate vocabulary semantic vectors to achieve more accurate vocabulary semantic analysis (Chen G., Xu, Hong, Wu, & Xiao, 2024). Commonly used word embedding models include Word2Vec, GloVe, and FastText (Borah, Barman, & Awekar, 2021). All three models use context information of words to capture the semantic relationship of words. Word2Vec optimizes the objective function to ensure that the distance between word vectors in similar contexts is close; FastText, based on Word2Vec, additionally captures structural information such as the internal character composition of words; GloVe represents semantics through the co-occurrence frequency of features in the entire corpus. Studies have shown that most word embedding models randomly initialize vectors, and the resulting semantic space is uncertain. Their default tokenizer often only performs simple word segmentation operations, and less work is done on screening feature combinations. On the same data, the word vectors generated by two trainings are different, and the semantic fields formed by the nearest neighbors of the words do not completely overlap (Kutuzov, Øvrelid, Szymanski, & Velldal, 2018; Rettenmeier, 2020). Therefore, reducing the uncertainty of word vectors is one of the key points of word embedding. For example, studies have shown that using a custom corpus can significantly improve the effect of text mining (Ercan & Cicekli, 2016; Nguyen, Billingsley, Du, & Johnson, 2015). In addition, N-gram Categories (i.e., phrases consisting of multiple words) show better performance in text classification (Kruczek, Kruczek, & Kuta, 2020).

Through the word embedding model, various text forms such as sentences, paragraphs, and documents can be represented as vectors, thereby realizing the combination with machine learning methods. One of the text clustering methods is to cluster text vectors into sentences, paragraphs, or documents through clustering methods such as K-means (Ji, Liu, Peng, & Kong, 2024). In addition to K-means, other commonly used algorithms for text clustering include Agglomerative Clustering (Enguix, Carrascosa, & Rincon, 2024), HDBSCAN (Inje, Nagwanshi, & Rambola, 2024), etc. Their clustering performance varies in different scenarios.

Text multi- feature extraction and seletion

High-frequency words have achieved rich application results in fields such as text topic classification and recognition (Qaiser & Ali, 2018; Tseng, Lin, & Lin, 2007). For example, the Vector Space Model (VSM) mainly uses word frequency to represent feature vectors and derives indicators such as TF-IDF, which is the most widely used (Choi, Oh, Choi, & Yoon, 2018). In addition, from the perspective of statistical features, the following two categories of text feature words are of particular concern. The first is to examine the ability of feature words to represent the characteristics of the technical domain from a global perspective, measure the

core influence and representativeness of feature words in the domain, and extract appropriate feature words with domain characteristics. The domain characteristic indicators used are mostly selected based on metrological and statistical features, mainly including TF-IDF (Chawla, Kaur, & Aggarwal, 2023), information gain (Yu, Ju, & Shang, 2022), Gini coefficient (Mengle & Goharian, 2009), and Comprehensively Measure Feature Selection (CMFS) (Yang, Liu, Zhu, Liu, & Zhang, 2012). They are mostly based on one kind of feature, among which CMFS integrates the comprehensive measurement of domain characteristics within and outside the class and has relatively good domain representation. The second is multi-feature, with more attention paid to technical interdisciplinary features (Yao, Wang, Wu, Xu, & Zhang, 2023) and technical breakthrough features (Jia et al., 2021; Liu Yahui, Xu Haiyun, Wu Huawei, Liu Chunjiang, & Wang Haiyan, 2023). Their effective identification methods are mostly achieved through relevant quantitative measures. Among them, interdisciplinary feature indicators mainly include Citation Outside Category index (COC) (Porter & Chubin, 1985), Weighted Citation Outside Category index (WCOC) (K. Chen & Chiung-fang, 2004), Shannon-Wiener Index (SWI) (Shannon, 1948), Brillouin's Index (BI) (Chang & Huang, 2012) and Terms Interdisciplinarity index (TI) (Xu, Guo, Yue, Ru, & Fang, 2016). Their main idea is to measure the degree of cross-integration between features. For example, the TI index considers cross-domain features and influence. So, its scalability is comprehensively good. Breakthrough features often have the characteristics of novelty, foresight, uncertainty, and nonlinearity. Scholars often start with the attributes of the technology itself or combine complex network methods for identification. For example, the identification method based on word frequency growth rate (Feng, Wu, & Mo, 2020), the identification method based on TRIZ theory (Vicente-Gomila, Artacho-Ramirez, Ting, & Porter, 2021), and the burst monitoring algorithm proposed by Kleinberg (Kleinberg, 2002). Among them, the Kleinberg burst detection algorithm is widely recognized by the academic community.

In the research on multi-feature technology topics mining, there is a significant impact on text mining results by feature selection (Büyükkeçeci & Okur, 2023). The results obtained by using features of different indicators and methods may be completely different(Zhang, Sun, Chinchilla-Rodríguez, Chen, & Huang, 2018). Scholars usually conduct comparative analysis of features from two major perspectives. First, from the information perspective, by comparing the differences in information content, richness, and synergy between different features, the similarity between different features is obtained based on indicators such as information entropy, mutual information, and information gain(Wang, Lu, & Tai, 2015), and the feature weights are assigned(Prabowo & Thelwall, 2006). The second is to explore the intrinsic connections between different features at the semantic level from a semantic perspective, thereby achieving topic clustering(Zhao, Guo, & Wu, 2024), feature fusion(Tien, Le, Tomohiro, & Tatsuya, 2019), etc. By analyzing the differences between different features, it is helpful to select appropriate features and apply them to text mining tasks such as topic representation. However, in the current research on text multi-feature, the academic community focuses more on how to

integrate features, and less on the selection of features combination and the influence of their mutual influence.

The word embedding model can convert the text feature words into vectors. It is one of the important steps in text clustering and is widely used in patent text analysis. Existing research starts from the perspective of multiple features, such as high frequency, field characteristics, technical intersections and technological breakthroughs. However, the default word embedding model is often implemented through simple tokenizer, lacks feature selection, and the certainty of the model needs to be improved. Scholars mostly compare and analyze different features from the information and semantic levels and rarely select multiple feature combinations. Therefore, existing research is insufficient in exploring the invisible relationships between different features and has not fully explained the complementary effects and coupling relationships of different features. In terms of the application of multiple features, it focuses on feature fusion but lacks feature selection methods. Therefore, explaining the specific complementary effects and coupling relationships between features, providing a basis for feature selection, and improving the certainty of the word embedding model is a key issue in improving the text clustering effect.

Methodology

Research Framework

Since patent documents are effective carriers of a large amount of world science and technology information, this study conducted research on patent texts. In view of the characteristics of technical themes, four types of patent text feature words, namely, text domain feature keywords (CMFS), technical interdisciplinary keywords (TI), technical breakthrough keywords (KB) and high-frequency keywords (HF), are selected as research objects. To fully explain the problem of complementary effects and coupling relationships between different features, based on the characteristics of patent text with strong technicality, obvious interdisciplinary features, and fast information changes, we combined the semantic and information levels, and selected the similarity, difference, complementarity and synergy between different features as the analysis target. In view of the problem of missing feature selection methods, the Jaccard distance is selected to measure the similarity and difference between features, and the information entropy and mutual information theory are combined to measure the information difference and synergy between features. The information difference and change between different feature words are compared and analyzed, and a comprehensive indicator is designed to select feature combinations. Based on the feature word list, we used Word2Vec, GloVe, and FastText to implement word embedding, and apply K-means, Agglomerative Clustering, and HDBSCAN algorithms to cluster the patent texts. By calculating the silhouette coefficient of each clustering result, we analyze the impact of different text feature combinations on text clustering, thereby verifying the effectiveness and feasibility of this method. The research method framework is shown in Figure 1.

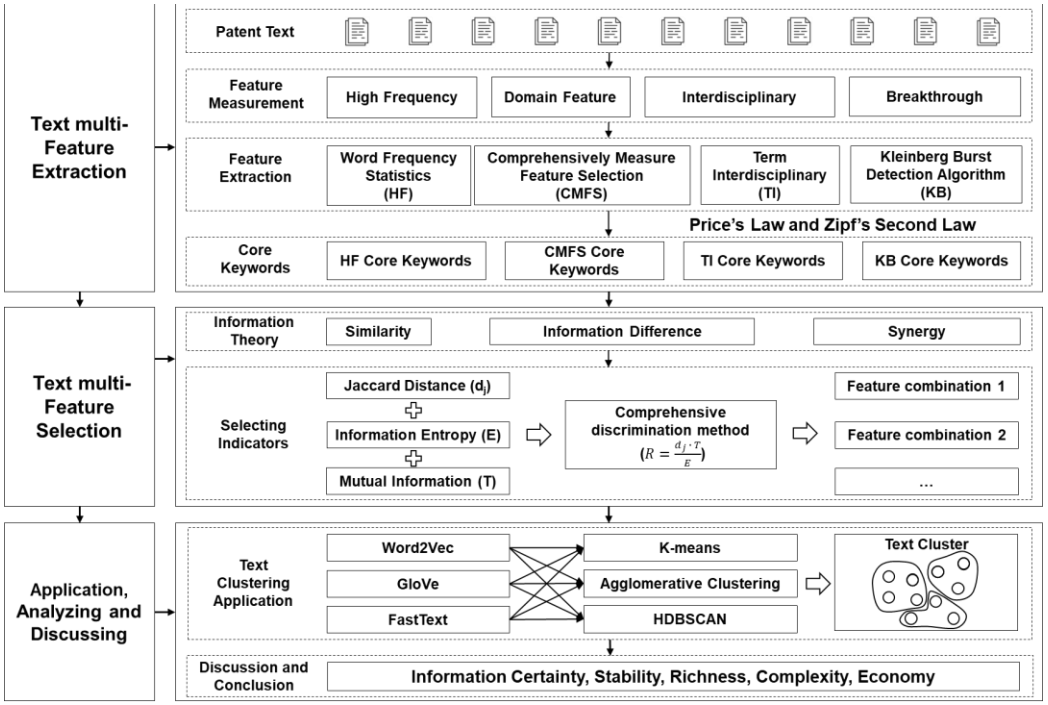


Figure 1. The research framework of clustering method by multi-text feature combination.

Data source and tokenizers

We took the field of graphene sensing technology as an example to carry out experimental research, extracting text domain feature keywords, technical interdisciplinary keywords, technical breakthrough keywords and high-frequency keywords, comparing the topic representation effects of the four types of characteristic keywords and their relationships. The reason for selecting graphene sensing technology for empirical research is that, firstly, there are strong interdisciplinary features in this technology, covering multiple technical fields such as materials, information, and biology; secondly, the breakthrough technology features and active technological innovations in this field are prominent, which have good practical significance for this study.

We selected the Derwent Innovation Index database as the data source. With the assistance of domain experts, we identified patent search strategies that are highly relevant to the topic of graphene sensing technology. The search date is October 31, 2022, and the search strategy is shown in Table 1. There were 974 items obtained after preliminary screening and elimination. Using the Derwent Data Analyzer (DDA) platform to perform NLP word segmentation processing based on the title and abstract text fields of 974 patent records, we obtained 20,036 original n-gram feature words (groups), where n ranges from 1 to 10. Then, the feature words were cleaned, using DDA's built-in stop words list, thesaurus, etc. The cleaning content includes removing common meaning stop words, formatting and grammatical terms of patent documents, DWPI description format abbreviations, compound name

specifications, British and American spelling specifications, etc. After cleaning, 16,604 feature words (groups) were obtained. Finally, manual cleaning is carried out. Personnel skilled in the field conduct manual interpretation, merge synonyms, and eliminate common feature words that are not closely related to substantive research, such as include, use, etc., as well as general experimental tool names, material names, etc. After cleaning, 7873 feature words (groups) were obtained as candidate feature items.

Table 1. The retrieval strategy for Grapheny Sensing Technology.

No.	*Search strategy*
# 1	TS=(sensor* or transducer* or (sensing same (element* or devic* or unit* or organ* or apparatus* or system*)) or (sense same organ*) or Photosensor* or microsensor* or chemosensor* or multisensory* or hypersensor*) database =Cderwent, Ederwent, Mderwent Timespan =2003-2022
# 2	TS=(graphene*) database =Cderwent, Ederwent, Mderwent Timespan =2003-2022
# 3	PN=(US*) database =Cderwent, Ederwent, Mderwent Timespan =2003-2022
# 4	#1 and #2 and #3 database =CDerwent, EDerwent, MDerwent Timespan =2003-2022

Text multi-feature extraction and evaluation

We selected word frequency, CMFS, TI, and KB as the feature extraction indicators, as shown in Table 2. Among them, further combining with the technical features of patent documents, when calculating TI, the IPC classification number is used to measure the technical intersection. At this time, the distribution degree *d* of the feature is the number of technical categories containing the feature, and *tf* is the frequency of the feature.

Table 2. Patent Text multi-feature measurement index.

Target	*Indicator*	*Methods*
High-Frequency Keywords	Word Frequency (Yan, Shuliang, Xiaochao, Yuhui, & Yafei, 2016)	$F = \sum_{i=0}^{N} f_i$ Measure the sum of word frequencies of the word in *N* documents.
Domain Feature Keywords	CMFS (Yang et al., 2012)	$CMFS_{avg}(t_k) = \sum_{i=1}^{\lvert C \rvert} \frac{P(t_k, c_i)(tf(t_k, c_i) + 1)^2}{(tf(t_k) + \lvert C \rvert)(tf(t, c_i) + \lvert V \rvert)}$

Technical Interdisciplinary Keywords	TI (Xu et al., 2016)	$tf(t_k, c_i)$represents the frequency of feature t_k in the *i*-th category c_i; $tf(t_k)$ represents the frequency of feature t_k in the entire training set; $tf(t, c_i)$ represents the sum of the frequencies of all features in category c_i; $\|C\|$ represents the number of categories; $\|V\|$ represents the number of features in the original vector space. $P(t_k, c_i)$ represents the frequency of feature t_k in the *i*-th category c_i as a percentage of the frequency of all categories $\|C\|$. $TI = d * \ln(tf)$ d is the degree of the feature words' distribution, and tf is the frequency of the feature words.
Technical Breakthrough Keywords	KB (Kleinberg, 2002)	$\sigma(i, r_t, d_t) = -\ln\left[\binom{d_t}{r_t} p_i^{r_t}(1-p_i)^{d_t - r_t}\right]$ $weight = \sum_{t=t_1}^{t_2} (\sigma(0, r_t, d_t) - \sigma(1, r_t, d_t))$ r_t is the frequency of target *i* at time *t*, d_t is the number of events in time *t*, and p_i is the frequency of the target in events.

Similarity and difference

The Jaccard similarity principle is used to calculate the similarities and differences between the comparison word lists. We used the complementary index of the Jaccard coefficient, the Jaccard distance d_j. The larger the Jaccard distance, the higher the difference between the sets. It is defined as follows(Jaccard, 1912):

$$\mathrm{d}_j(A,B) = 1 - \mathrm{Jaccard}(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (1)$$

To facilitate the quantification of the differences between multiple sets, all two sets that do not repeat are taken from the multiple sets, and their Jaccard coefficients are calculated respectively. Then, the average value of the Jaccard coefficients of all two sets is calculated, which is defined as Equation 2, where n is the number of sets.

$$\mathrm{d}_j(A,B,\ldots,N) = \frac{\sum \mathrm{d}_j(x,y)}{C(n,\ 2)} \quad (x,y \in (A,B,\ldots,N), x \neq y) \quad (2)$$

Information difference and synergy

We also introduced the concept of information entropy and mutual information measurement method, places the feature words in the context of sentences, and quantitatively detects the amount of information revealed by the feature words list and the degree of overlap and fusion between them.

The uncertainty of information corresponds to information entropy. Shannon borrowed the concept of thermodynamics and defined the mathematical expectation of self-information as "information entropy" to measure the amount of information. Combined with linguistic improvements, the probability P_x in the formula is expressed as the relative frequency of a certain feature (that is, the ratio of the feature frequency to the total number of all feature frequencies), and the information entropy calculation formula for measuring the amount of information is obtained as follows (Shannon, 1948):

$$E_x = -\sum_x P_x \lg(P_x) \quad (3)$$

The smaller the information entropy, the more information the text information is concentrated on one or some features; the larger the information entropy, the more information it carries, and the richer or more variable its features are, and the greater its uncertainty. For two-dimensional events, the information entropy E is as follows: Where P_{xy} is the joint probability distribution of event x and event y.

$$E_{xy} = -\sum_x \sum_y P_{xy}\, lg\left(P_{xy}\right) \quad (4)$$

Mutual information is the amount of information about another random variable contained in a random variable, that is, the uncertainty of a random variable reduced by knowing another random variable. It can measure the uncertainty transfer degree between subsystems, that is, the synergy relationship. Abramson (Abramson, 1963) used the mutual information measure of subsystem variables to define the mutual information transfer degree of two interacting subsystems and three interacting subsystems as follows:

$$T_{xy} = E_x + E_y - E_{xy} \quad (5)$$

$$T_{xyz} = E_x + E_y + E_z - E_{xy} - E_{xz} - E_{yz} + E_{xyz} \quad (6)$$

Based on the mutual information measurement theory and application research, we constructed four types of vocabulary synergy. According to the chain rule of mutual information, as follows:

$$T(x_1, x_2, \cdots, x_n; y) = E(x_1, x_2, \cdots, x_n) - E(x_1, x_2, \cdots, x_n|y) \quad (7)$$

Then, specifically for the four vocabularies of HF, CMFS, TI, and KB, their synergy T_{hctk} can be defined as:

$$T_{hctk} = E_h + E_c + E_t + E_k - E_{hc} - E_{ht} - E_{hk} - E_{ct} - E_{ck} - E_{tk} + E_{hct} + E_{hck} + E_{htk} + E_{ctk} - E_{hctk} \quad (8)$$

As a quantitative indicator, T_{hctk} measures the uncertainty of the interaction between the four types of vocabularies, thereby reflecting the degree of information fusion and interaction between the four types of features. T_{hctk} is a positive indicator. The larger the T_{hctk} value, the stronger the interaction and synergy of the four features.

Comprehensive Evaluation

Mutual information is often used for feature selection, and especially has good performance in feature dimensionality reduction(Gandhi & Prabhune, 2017). However, mutual information is difficult to filter out information redundancy. The academic community often combines information entropy to maximize mutual information and minimize information entropy to comprehensively select features, while reducing feature dimensionality, filtering out information redundancy(Liu & Wen, 2023). Therefore, this study calculated the ratio of the two to balance information entropy and mutual information. On this basis, feature combinations with good complementarity and low repetition rate are preferred. Therefore, the Jaccard distance is added to the numerator of the ratio fraction, and the final selection measurement index of the feature combination is obtained as follows.

$$R = \frac{d_j \cdot T}{E} \quad (9)$$

The larger the R is, the more likely it is that the feature combination has higher information certainty, higher information synergy, and better complementarity among the features within this combination compared to any other combination.

Word embedding and text clustering

We used Word2Vec, GloVe, and FastText models to implement word embedding for different feature combinations. Since most pre-trained models would converge after the word vector dimension reaches 300, this study set the word vector dimension to 300 and uses the weighted average of all word vectors in the document (Equation 10) to represent the document. We applied three algorithms: K-means, Hierarchical Clustering (Agglomerative Clustering), and HDBSCAN to cluster patent texts. Through combination, 9 different text clustering models can be obtained. By calculating the silhouette coefficient of each clustering result, the influence of different feature combinations on the text clustering effect is measured. The silhouette coefficient is a clustering performance evaluation index that objectively reflects the outline clarity of each clustering cluster. Its calculation formula is shown in Equation 11(Bagirov, Aliguliyev, & Sultanova, 2023).

$$v_W = \sum v_i * p_i \quad (10)$$

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \quad (11)$$

Among them, v_W represents the vector of document W, v_i represents the vector of feature i, p_i is the frequency of feature i in document W, a_i is the average distance

between each data point i and all other points in the same cluster, and b_i is the average distance between each data point i and all points in the nearest non-self-cluster. The value range of the silhouette coefficient s_i is [-1, 1]. Close to 1 means that the data point is very similar to other points in its own cluster and has obvious differences from data points in other clusters, and the clustering effect is good; while close to 0 means that the data point is on the boundary of two clusters, and the clustering effect is average; close to -1 means that the data point may be mistakenly assigned to the wrong cluster, and the clustering effect is poor.

Results

Text multi-feature extraction result

For 7873 feature words (groups), the Comprehensively Measure Feature Selection (CMFS), Term Interdisciplinary index (TI), Kleinberg burst detection algorithm, and word frequency statistics were used to measure four types of text features. We extracted four types of feature words through programming. According to the measurement results of the feature value of each feature word, the CMFS keywords, TI keywords, KB keywords, and HF keywords in the field of graphene sensing were obtained. Taking the top 20 words as an example, the results of four types of feature-word lists are shown in Table 3.

Next, it is necessary to determine the effective threshold for each feature value, in order to select the appropriate amounts of core keywords of HF, CMFS, TI, and KB respectively, and form a thesaurus with effective topic representation meaning. For threshold determination, this study applied the ideas of Price's law and Zipf's second law. Price's formula was first used to determine highly cited literature and then determine the core authors in a certain research field. It is a scientific method for selecting thresholds and has gradually been applied by scholars in different research fields. Here, we used Price's formula to determine the threshold for core keywords, with the independent variable N_{max} representing the maximum value of the keyword's frequency, TI and KB, to obtain the core keywords threshold for each word list. The calculation formula is as follows (Price, 1963):

Table 3. The extraction results of keywords of HF, CMFS, TI and KB in graphene sensing field (TOP 20).

NO.	HF Keywords	Frequency	CMFS Keywords	CMFS (×10-7)	TI Keywords	TI	KB Keywords	KB
1	method	1380	three dimensional image	2.28	sensor	639.32	method	72.23
2	layer	914	sensor mounting	2.18	method	571.16	surface	50.32
3	sensor	899	conductive membrane	2.13	detecting	489.49	sensor	49.63
4	surface	804	manganese content	2.12	patient	474.78	device	42.80
5	patient	802	toilet seat	2.10	device	467.13	chemistry	39.86
6	device	791	spring structure	2.09	system	442.08	patient	38.28
7	detecting	746	transducer assembly	2.08	surface	441.51	substrate	37.17
8	substrate	671	ultrasound transducer element	2.05	material	425.27	electrode	34.78
9	electrode	660	nano-cone structure polypyrrole	2.05	substrate	410.05	layer	34.57
10	analyte	594	carbon particle	2.02	graphene	394.30	graphene	32.80
11	system	553	enzyme solution	2.01	chemistry	368.09	glucose	29.40
12	sample	531	temperature difference	2.01	layer	361.35	system	27.60
13	glucose	499	carbon nanowalls-based breath sensor	2.00	electrode	318.12	concentration	23.56
14	protein	475	sol-gel silicon film	2.00	solution	310.66	protein	22.03
15	solution	442	lithium ion battery	1.99	polymers	310.57	polymers	21.83
16	material	435	body sensor network	1.99	glucose	291.99	electrodes	19.08
17	graphene	431	elastic container	1.99	concentration	282.80	metal	13.74
18	antibody	412	cervical cancer	1.98	electrodes	270.50	working electrode	12.63
19	binding	397	carbon quantum dot	1.98	polymer	262.96	data	11.90
20	nanoparticles	381	metal hydroxide quantum dots	1.97	protein	234.21	sensitivity	10.35

$$M = 0.749\sqrt{N_{max}} \quad (12)$$

Due to the significant scale difference between the CMFS feature values and the other three types of feature values, the sensitivity of the Price formula in distinguishing the core words of CMFS is poor. So, we applied Zipf's second law to calculate the threshold of CMFS core keywords. The calculation formula is as follows (Donohue, 1973), where I is set to the maximum value of CFMS.

$$T = \frac{1}{2}(-1 + \sqrt{1 + 8 * I}) \quad (13)$$

According to the calculation results of the core keywords thresholds of each words list, the results of four types of core keywords are shown in Table 4.

Table 4. The threshold and number of core keywords of HF, CMFS, TI and KB.

Keywords	*HF*	*CMFS*	*TI*	*KB*
Core Keyword Threshold	27.82	1.69	18.94	6.37
Number of Core Keywords	305	185	608	29

Text multi-feature combination discrimination results

Similarity and difference

The overlap and Jaccard distance between the CMFS, TI, KB core keywords and the HF core keywords are calculated respectively, as shown in Table 5.

Table 5. The differences between core keywords of the CMFS, TI, KB and HF.

Feature	*CMFS vs. HF*	*TI vs. HF*	*KB vs. HF*
Overlaps Number	11	285	29
Overlaps Rate	$\frac{CMFS \cap HF}{HF} = 3.60\%$	$\frac{TI \cap HF}{HF} = 93.44\%$	$\frac{KB \cap HF}{HF} = 9.51\%$
Number of Core Keywords	$\frac{CMFS \cap HF}{CMFS} = 5.95\%$	$\frac{TI \cap HF}{TI} = 46.88\%$	$\frac{KB \cap HF}{KB} = 100\%$

The results show that: (1) The overlap rate between the CMFS core keywords and the HF core keywords is the lowest, and the Jaccard distance is the largest, that is, the difference between the CMFS and HF keywords is the largest. This shows that in terms of text feature representation, CMFS core keywords can reveal important thematic features that HF cannot reflect and may play a complementary role for the HF vocabulary. (2) The overlap rate between the TI core keywords and the HF core keywords is high, and the Jaccard distance is relatively small. The technical interdisciplinary has the characteristics of a wide range, but not completely overlap.

This shows that the TI core keywords can effectively identify some low-frequency words with technical interdisciplinary characteristics. (3) The KB core keywords have the least number of words, and all of them belong to HF core keywords, which is consistent with the explosive growth of technology breakthrough in a short period of time. However, its supplementary role in the HF core keywords is of little significance.

Furthermore, the core keywords of CMFS, TI, and KB are compared in pairs, and their overlap rate and Jaccard distance are calculated. As shown in Table 6.

Table 6. The differences between core keywords of CMFS, TI, KB.

Feature	*CMFS vs. TI*	*TI vs. KB*	*CMFS vs. KB*
Overlaps Number	7	29	0
Overlaps Rate	$\frac{\text{CMFS} \cap \text{TI}}{\text{CMFS}} = 3.78\%$	$\frac{\text{TI} \cap \text{KB}}{\text{TI}} = 4.77\%$	$\frac{\text{CMFS} \cap \text{KB}}{\text{CMFS}} = 0\%$
Number of Core Keywords	$\frac{\text{CMFS} \cap \text{TI}}{\text{TI}} = 1.15\%$	$\frac{\text{TI} \cap \text{KB}}{\text{KB}} = 100\%$	$\frac{\text{CMFS} \cap \text{KB}}{\text{KB}} = 0\%$

The results show that: (1) All KB core keywords overlap with the TI core keywords, which displays that most technical breakthrough words may also have technical intersection attributes. (2) The CMFS core keywords do not overlap with the KB core keywords list at all, and the overlap rate with the TI core keywords is very low. (3) The overlap rate of TI core keywords and CMFS is extremely low. Overall, three types of core keywords show good text feature complementarity, especially the CMFS keywords and the TI keywords.

Information difference and synergy

To explore the features at the sentence level, this study segmented the patent document text into sentences. We selected sentences containing HF core words, CMFS core words, TI core words, and KB core words, and classified them into four types of patent text sets. There are 11 different combinations of the four types of features, resulting in 11 types of text sets. The information entropy and mutual information of each text set are calculated as shown in Table 7. Figure 2 intuitively presents the changes in information entropy and mutual information of texts with different feature words.

Table 7. The Information Entropy and Mutual Information of text sets of HF, CMFS, TI and KB.

NO.	*Feature*	*Extraction Rules*	*Sentences Number*	*Information Entropy*	*Mutual Information*
#1	HF	HF core keyword appears in the sentence.	9502	0.072	-
#2	CMFS	CMFS core keyword appears in the sentence.	1215	0.102	-
#3	TI	TI core keyword appears in the sentence.	9930	0.059	-
#4	KB	KB core keyword appears in the sentence.	6182	0.146	-
#5	HF + CMFS	HF, CMFS core keyword appears in the sentence together.	1011	0.092	0.082
#6	HF + TI	HF, TI core keyword appears in the sentence together.	9475	0.073	0.058
#7	HF + KB	HF, KB core keyword appears in the sentence together.	6182	0.146	0.072
#8	CMFS + TI	CMFS, TI core keyword appears in the sentence together.	1041	0.094	0.068
#9	CMFS + KB	CMFS, KB core keyword appears in the sentence together.	614	0.067	0.181
#10	TI + KB	TI, KB core keyword appears in the sentence together.	6182	0.146	0.059
#11	HF + CMFS + TI	HF, CMFS and TI core keyword appears in the sentence together.	1006	0.092	0.067
#12	HF + CMFS + KB	HF, CMFS and KB core keyword appears in the sentence together.	614	0.067	0.082
#13	HF + TI + KB	HF, TI and KB core keyword appears in the sentence together.	6182	0.146	0.058
#14	CMFS + TI + KB	CMFS, TI and KB core keyword appears in the sentence together.	614	0.067	0.068
#15	HF + CMFS + TI + KB	HF, CMFS, TI and KB core keyword appears in the sentence together.	614	0.067	0.067

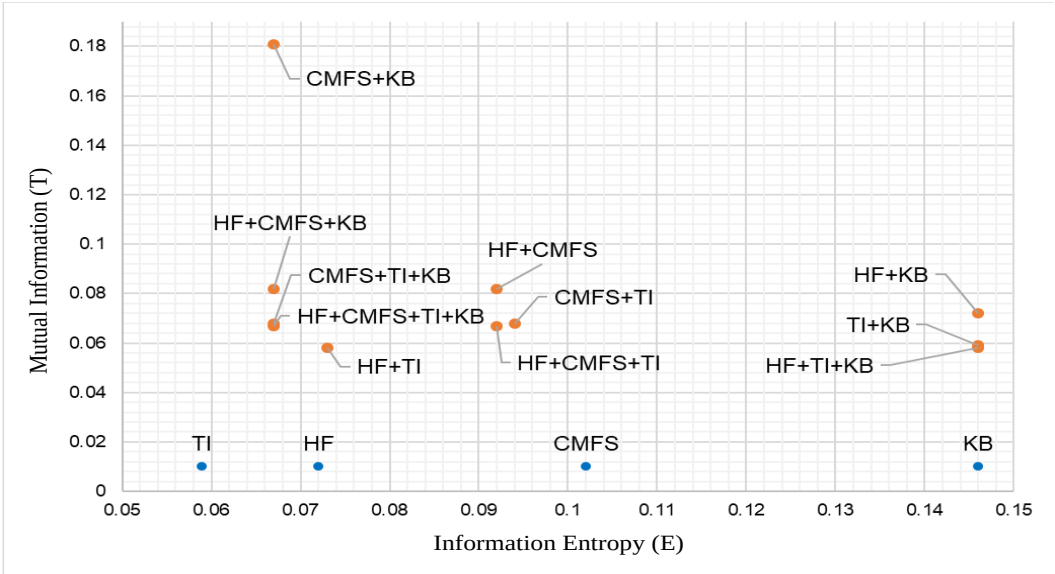


Figure 2. The Information Entropy and Mutual Information of HF, CMFS, TI, and KB texts.

The comparative analysis results of the Information entropy suggest that: (1) The descending order of the number of sentences containing the four types of core keywords is: TI > HF > KB > CMFS. The order of information uncertainty from high to low is KB > CMFS > HF > TI (#1 to #4). (2) The information entropy of TI text (#3) is lowest, while the KB text (#4) is highest, showing that the text feature concentration of the technology interdisciplinary is the highest, and the text feature complexity of the technology breakthrough is the highest. (3) The topic complexity of HF+KB texts is higher than HF texts, while HF+CMFS slightly increases topic complexity, and HF+TI has a smaller change (#1, #5, #6, #7). (4) The information entropy of CMFS text and KB text is relatively high (#2, #4). When either of them is combined with HF or TI features, it can improve the information entropy of the original HF or TI features (#5, #7, #8, #10). When CMFS and KB features appear at the same time (#9), the information entropy of the text decreases significantly. When CMFS+KB features are combined with other features at the same time, the information entropy of the text decreases significantly relative to other features (#12, #14). Therefore, the information uncertainty of CMFS and KB features is high individually, but when they are used simultaneously, the uncertainty is greatly reduced. (5) The four types of texts, namely KB, HF+KB, TI+KB, and HF+TI+KB features texts, have the same number of sentences and information entropy (#4, #7, #10, and #13), which shows that KB features are always accompanied by HF and TI features. This reveals to a certain extent that technological breakthroughs often occur when the development of technology accumulates to a certain extent and intersects. (6) CMFS features significantly reduce the information richness of the HF+TI+KB text (#11, #12, #13, #14, #15).

The comparative analysis results of the mutual information suggest that: (1) the synergy of the CMFS+KB (#9) text is the highest, indicating that there is a certain

information sharing between CMFS and KB features, that is, when one type of feature appears in a sentence, the certainty of the other type of feature will further increase. (2) The mutual information of HF+TI, TI+KB, and HF+TI+KB is relatively small (#6, #10, #13), indicating that the interactivity and synergy of HF, TI, and KB are relatively low. (3) The mutual information of the TI+KB text is almost the smallest and the information entropy is the largest (#10), but the information entropy of the TI text is the lowest (#3) while the information entropy of the KB text is the highest (#4). So, the synergy of the TI and KB features is relatively low. Combined with the quantitative characteristics of TI and KB core keywords (Table 6), the KB core keywords are less in number than the TI core keywords, but the information content is richer. So, the information gain effect of the TI core keywords on the KB core keywords is relatively small.

Taken together, when only a certain type of feature needs to be extracted, HF features and TI features involve rich sentences and the information certainty is highest, so they can be preferred as basic word lists. KB features contain a large amount of information but high levels of uncertainties. Although there are many sentences involved, the number of core words is small. Therefore, KB features can be used in combination with HF features and TI features to enhance the information richness of the latter two. CMFS features can enhance the stability of HF features and TI features. CMFS features and KB features are highly synergistic in text, and their combined use can significantly reduce topic uncertainty. From the perspective of computational complexity, when giving priority to information richness, the most economical choice is HF+KB core words. When giving priority to information certainty, the most economical choice is CMFS+KB core words. When considering a compromise between the two, HF+CMFS+KB core words are a suitable choice

Comprehensive discrimination method

To comprehensively balance the differences, information certainty and information synergy between different features in the feature combination, the comprehensive discrimination index *R* of each feature combination is calculated. The results are shown in Table 8.

Table 8. The Comprehensive discrimination index of each feature combination.

NO.	*Feature*	*Information Entropy*	*Mutual Information*	*Jaccard distance*	*Comprehensive discrimination*
#5	HF + CMFS	0.092	0.082	0.977	0.871
#6	HF + TI	0.073	0.058	0.546	0.434
#7	HF + KB	0.146	0.072	0.905	0.446
#8	CMFS + TI	0.094	0.068	0.99	0.716
#9	CMFS + KB	0.067	0.181	1	**2.701**
#10	TI + KB	0.146	0.059	0.95	0.384
#11	HF + CMFS + TI	0.092	0.067	0.838	0.610
#12	HF + CMFS + KB	0.067	0.082	0.961	**1.176**
#13	HF + TI + KB	0.146	0.058	0.800	0.318
#14	CMFS + TI + KB	0.067	0.068	0.980	**0.995**
#15	HF + CMFS + TI + KB	0.067	0.067	0.895	0.895

Based on Table 8, the three combinations with the highest comprehensive discrimination indexes are selected, namely, CMFS + KB, HF + CMFS + KB, and CMFS + TI + KB. The number of KB features used alone is small, and it is difficult to obtain valuable information. While CMFS+KB significantly reduces topic uncertainty and provides information supplements for the scarce breakthrough features, the CMFS+KB focuses on the characterization of technology breakthrough features. Since the repetition rate between the KB and the TI and HF is high, the HF+CMFS+KB focuses on the characterization of technology domain features, while the CMFS+TI+KB focuses more on the characterization of interdisciplinary. So far, we have selected three sets of feature combinations with better comprehensive representation capabilities.

Patent text clustering based on multi-feature combination

To further explore the impact of different feature combinations on text clustering, we applied 9 different text clustering models to the three sets of features combinations and calculated the silhouette coefficient of each clustering algorithm as shown in Table 9. This patent dataset covers 8 major IPC categories, so in the algorithm that requires the input of the number of clusters, the default number of clusters is 8. For easy comparison, this study also applied the clustering model to the HF features and used principal component analysis (PCA) to reduce the dimension of each text to 2 dimensions and visualized it as shown in Figure 3.

Table 9. Silhouette coefficients of text clustering based on different feature combinations.

Feature Combinations	*Word Embedding*	*K-means*	*Agglomerative Clustering*	*HDBSCAN*	*Mean*
HF	GloVe	0.066	0.038	0.019	0.041
	Word2Vec	0.161	0.141	-0.059	0.081
	FastText	0.204	0.158	0.089	0.150
	Mean	0.144	0.112	0.016	Overall mean 0.091
HF + CMFS + KB	GloVe	0.071	0.035	0.009	0.038
	Word2Vec	0.166	0.132	0.001	0.100
	FastText	0.203	0.164	**0.107**	**0.158**
	Mean	0.147	0.110	**0.039**	Overall mean **0.099**
CMFS + TI + KB	GloVe	0.064	0.022	0.018	0.035
	Word2Vec	0.160	0.135	0.008	**0.101**
	FastText	0.176	0.152	0.092	0.140
	Mean	0.116	0.082	0.016	Overall mean 0.071
CMFS + KB	GloVe	0.131	0.083	0.010	**0.075**
	Word2Vec	0.178	0.123	-0.133	0.056
	FastText	**0.216**	**0.170**	0.064	0.150
	Mean	**0.175**	**0.125**	-0.020	Overall mean 0.094

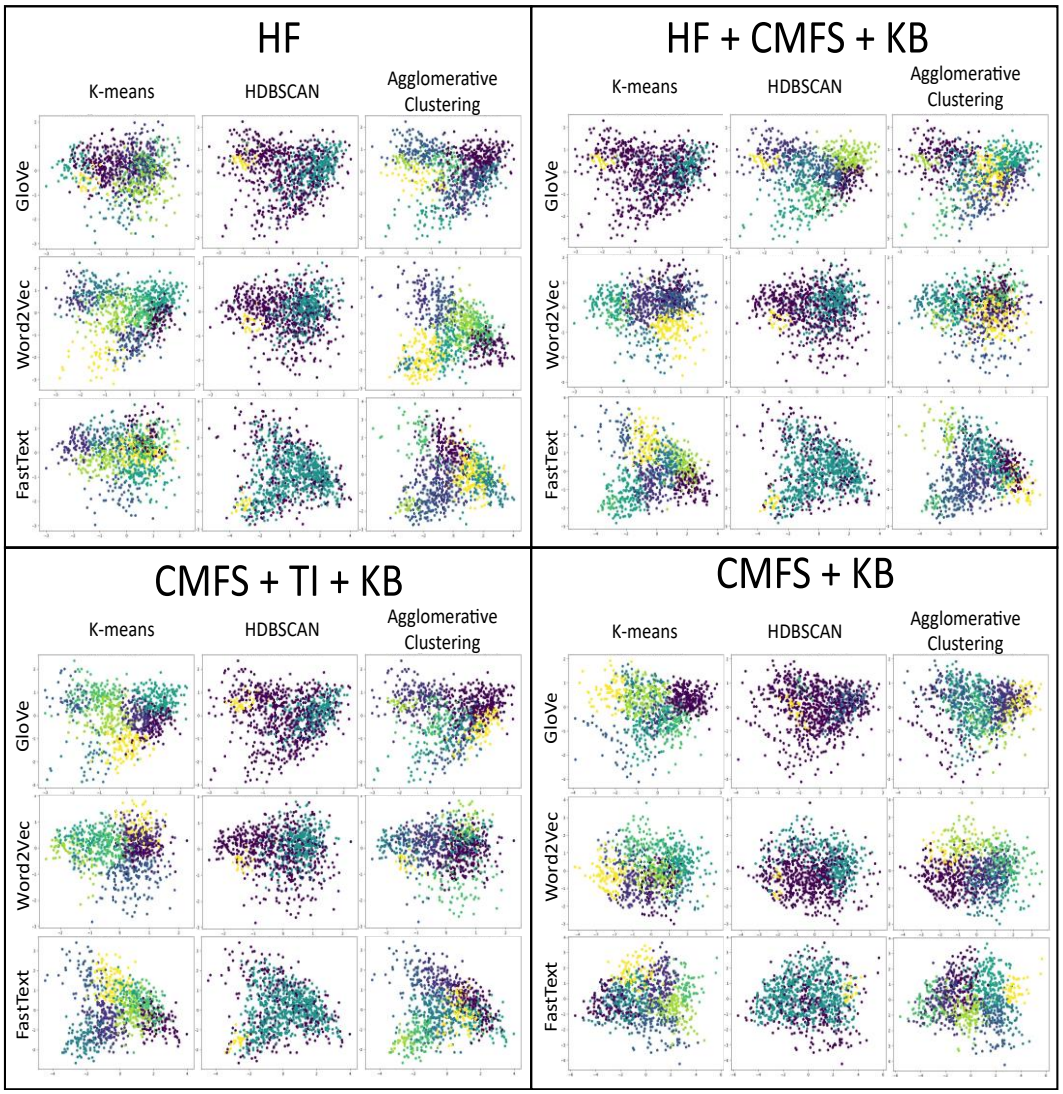


Figure 3. Clustering scatter plots of different feature combinations by PCA.

As shown in Table 9, overall, the overall average silhouette coefficient of feature combinations HF + CMFS + KB and CMFS + KB is greater than HF in the 9 models, which effectively improves the patent text clustering effect. For different word embedding models, HF + CMFS + KB performs best in FastText, CMFS + TI + KB performs best in Word2Vec, and CMFS + KB performs best in GloVe. For different clustering algorithms, HF + CMFS + KB performs best in HDBSCAN, and CMFS + KB performs best in K-means and Agglomerative Clustering.

Combining the comparative analysis results of the differences, information certainty and synergy of different feature combinations, the information certainty, synergy and stability of CMFS + KB feature combination is high. It achieved good results in multiple models and algorithms, especially when used with FsatText+K-means. But for Word2Vec and HDBSCAN, the advantage is not obvious; HF + CMFS + KB slightly improves the information certainty compared to CMFS + KB, but the

synergy decreases significantly. It is more suitable for the specific HDBSCAN+FastText model.

Discussion

Based on the above research and analysis, the main discussions are as follows:

(1) From the perspective of feature morphological differences, the complementarity of three feature words of CMFS, TI, and HF is good in general. Among them, the CMFS features have the best supplementary effect on the HF features and can supplement some features with domain characteristics but without high frequency. The CMFS features do not overlap with the KB features at all, and the overlap between other features is relatively low. The overlap between TI features and the KB or HF features is relatively high, but TI features can effectively identify some non-high frequency and interdisciplinary characteristics, which has a certain supplementary significance for the HF features. KB features completely overlap with HF and TI features, of which have both technical breakthrough and technical interdisciplinary characteristics. It has the weakest supplementary significance for the HF features and can only realize the selection of features with technical breakthrough characteristics from HF features.

(2) From the perspective of text information, the significance of CMFS features in revealing text topic characteristics is stronger than other features. The information concentration of TI features is the highest; the complexity of KB features is the highest. The CMFS and KB features are strongly uncertainty, but when they appear at the same time, the uncertainty of information is greatly reduced. Although HF+KB features are rich in information, they have a high repetition rate and are difficult to supplement the HF features.

(3) From the perspective of the synergy of feature texts, the synergy of CMFS and KB features is highest, and information certainly is better. There is good information sharing between CMFS+KB. For other feature combinations, the mutual information performance is relatively low, but the information certainly exhibits a variety. Among them, HF+CMFS+KB feature combination is a compromise between the synergy and information certainty.

(4) Considering the complementarity, information certainty, and synergy, the combined features of CMFS + KB, HF + CMFS + KB, and CMFS + TI + KB are better in representation.

(5) From the application effect of text clustering, the CMFS + KB is widely applicable in text clustering tasks of various word embedding and clustering algorithms. When used with FsatText + K-Means, the text clustering effect is the best. For the specific FastText + HDBSCAN model, FastTex incorporates word substring information, while HDBSCAN constructs a distance matrix and a directed weighted graph, resulting in higher computation complexity of the model. In this case, HF + CMFS + KB performs better.

In general, compared with the use of HF features alone, the combination of text multi-feature effectively improves the clustering effects. Among them, the CMFS+KB feature combination greatly improves the information certainty, resulting in a good stability of the clustering model to a certain extent, and helps to

identify the target topic more accurately and quickly. For the models with relatively high complexity, the HF + CMFS + KB feature combination may be effective. This study provides a certain basis and support for the selection and combination of vocabulary in text clustering.

Conclusion

Studying the limitation of insufficient quantitative measurement of the meaning of text multi-features in current text mining, we extracted four types of text features from patent texts, including domain feature, technical interdisciplinary, technical breakthrough and high frequency. Combined with the characteristics of patent texts, such as strong technicality, obvious interdisciplinary, and fast information changes, the similarity, difference, complementarity, and synergy between different thematic features are selected as analysis targets. Employing the Jaccard distance combined with the information entropy and mutual information theory, a method for comparative analysis of information differences and changes between different features is designed. A comprehensive discrimination index for feature combination selection is proposed. Using the Jaccard distance, information entropy, and mutual information index, a quantitative comprehensive measurement of four text features representation meaning is carried out. Taking the field of graphene sensor technology as an example, the similarity, difference, synergy, and complementarity of the four text features recognition are compared. Through comprehensive discrimination indicators, three representative feature combinations are selected, and they are applied in combination with a variety of word embedding models and clustering algorithms. The research results show that compared with simple high-frequency features, text multi-features can effectively play a complementary role from different perspectives. Selecting different feature combinations for use can reduce the uncertainty of text information, enhance the richness of text information, and improve the stability of text information to a certain extent. In addition, empirical analysis based on graphene sensing technology also provides optimization inspiration for parameter fitting and feature training of various language or topic models, thereby improving the recognition accuracy of unique feature technologies. This method can meet the needs of different analysis purposes and application scenarios in actual applications and help improve the efficiency and accuracy of technological innovation research.

Limitations

This study also has certain limitations. Since the research focus is on the quantitative and selection measurement of text multi-features, the extraction methods for the features are not rich enough, which may affect the richness of the text features. In the future, we can further enrich the measurement and extraction of different text features, explore more nonlinear relationships between different text features.

Acknowledgments

This contribution is the outcome of the projects, “Research on Multiple Relationship Fusion Methods for Identification and Prediction of Technological Innovation Paths” (No.18BTQ067) supported by the National Social Science Fund of China, “Early Recognition Method of Transformative Scientific and Technological Innovation Topics based on Weak Signal Temporal Network Evolution analysis” (No.72274113) supported by the National Natural Science Foundation of China, “Youth Innovation Promotion Association (2022173)” supported by Chinese Academy of Sciences(CAS), and the Taishan Scholar Foundation of Shandong province of China (tsqn202103069 and tsqn202103070).

References

Abramson, N. (1963). Information theory and coding (First Edition.). New York: McGraw Hill.

Bagirov, A. M., Aliguliyev, R. M., & Sultanova, N. (2023). Finding compact and well-separated clusters: Clustering using silhouette coefficients. Pattern Recognition, 135, 109144.

Borah, A., Barman, M. P., & Awekar, A. (2021). Are word embedding methods stable and should we care about it? Proceedings of the 32nd ACM Conference on Hypertext and Social Media, HT ’21 (pp. 45–55). New York, NY, USA: Association for Computing Machinery. Retrieved January 12, 2025, from https://doi.org/10.1145/3465336.3475098

Büyükkeçeci, M., & Okur, M. C. (2023). A comprehensive review of feature selection and feature selection stability in machine learning. Gazi University Journal of Science, 36(4), 1506–1520.

Chang, Y.-W., & Huang, M.-H. (2012). A study of the evolution of interdisciplinarity in library and information science: Using three bibliometric methods. Journal of the American Society for Information Science and Technology, 63(1), 22–33.

Chawla, S., Kaur, R., & Aggarwal, P. (2023). Text classification framework for short text based on TFIDF-FastText. Multimedia Tools and Applications, 82(26), 40167–40180.

Chen G., Xu Z., Hong S., Wu J., & Xiao L. (2024). A Study on the Stability of Semantic Representation of Entities in the Technology Domain-Comparison of Multiple Word Embedding Models. Journal of the China Society for Scientific and Technical Information, 43(12), 1440–1452.

Chen, K., & Chiung-fang, L. (2004). Disciplinary interflow of library and information science in taiwan. Journal of Library and Information Studies, 2.

Cheng Yong, Xu Dekuan, & Lv Xueqiang. (2019). Automatically grading text difficulty with multiple features. Data Analysis and Knowledge Discovery, 3(7), 103–112.

Chi, J., Ouyang, J., Li, C., Dong, X., Li, X., & Wang, X. (2019). Topic representation: Finding more representative words in topic models. Pattern Recognition Letters, 123, 53–60.

Choi, H., Oh, S., Choi, S., & Yoon, J. (2018). Innovation topic analysis of technology: The case of augmented reality patents. IEEE Access, 6, 16119–16137. Presented at the IEEE Access.

Donohue, J. C. (1973). Understanding scientific literatures: A bibliometric approach. The MIT Press, 28 Carleton Street, Cambridge, Massachusetts 02142 ($12.

Enguix, F., Carrascosa, C., & Rincon, J. (2024). Exploring federated learning tendencies using a semantic keyword clustering approach. Information, 15(7), 379.

Ercan, G., & Cicekli, I. (2016). Topic segmentation using word-level semantic relatedness functions. Journal of Information Science, 42(5), 597–608.

Feng, G., Wu, J., & Mo, X. (2020). Research on detection and verification of burst words with multiple measures. Library and Information Service, 64(11), 67–76.

Feng Guohe & Kong Yongxin. (2020). Subject hotspot research based onWord frequency analysis of time-weighted keywords. Journal of the China Society for Scientific and Technical Information, 39(1), 100–110.

Gandhi, S. S., & Prabhune, S. S. (2017). Overview of feature subset selection algorithm for high dimensional data. 2017 International Conference on Inventive Systems and Control (ICISC) (pp. 1–6). Presented at the 2017 International Conference on Inventive Systems and Control (ICISC). Retrieved January 9, 2025, from https://ieeexplore.ieee.org/document/8068599

Inje, B., Nagwanshi, K. K., & Rambola, R. K. (2024). An efficient document information retrieval using hybrid global search optimization algorithm with density based clustering technique. Cluster Computing, 27(1), 689–705.

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. New Phytologist, 11(2), 37–50.

Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. Journal of Memory and Language, 87, 38–58. San Diego: Academic Press Inc Elsevier Science.

Ji D., Liu Y., Peng R., & Kong H. (2024). K-means text clustering algorithm based on the center point of subject word vector. Computer Applications and Software, 41(10), 282–286, 318.

Jia, W., Xie, Y., Zhao, Y., Yao, K., Shi, H., & Chong, D. (2021). Research on disruptive technology recognition of China's electronic information and communication industry based on patent influence. Journal of Global Information Management (JGIM), 29(2), 148–165. IGI Global.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '02 (pp. 91–101). New York, NY, USA: Association for Computing Machinery. Retrieved January 14, 2024, from https://doi.org/10.1145/775047.775061

Kruczek, J., Kruczek, P., & Kuta, M. (2020). Are n-gram categories helpful in text classification? In V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, & J. Teixeira (Eds.), Computational Science – ICCS 2020 (Vol. 12138, pp. 524–537). Presented at the 20th Annual International Conference on Computational Science (ICCS), Cham: Springer International Publishing. Retrieved September 16, 2024, from https://link.springer.com/chapter/10.1007/978-3-030-50417-5_39

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018, June 13). Diachronic word embeddings and semantic shifts: A survey. arXiv. Retrieved January 13, 2025, from http://arxiv.org/abs/1806.03537

Li, X., Zhang, A., Li, C., Ouyang, J., & Cai, Y. (2018). Exploring coherent topics by topic modeling with term weighting. Information Processing & Management, 54(6), 1345–1358.

Liu Y., & Wen Y. (2023). Feature Selection Based on Maximizing Joint Mutual Information and Minimizing Joint Entropy. Advances in Applied Mathematics, 12, 1451.

Liu Yahui, Xu Haiyun, Wu Huawei, Liu Chunjiang, & Wang Haiyan. (2023). Scientific breakthrough topics identification in an early stage using multiple weak linkage fusion. Journal of the China Society for Scientific and Technical Information, 42(1), 19–30.

Mengle, S. S. R., & Goharian, N. (2009). Ambiguity measure feature-selection algorithm. Journal of the American Society for Information Science and Technology, 60(5), 1037–1050.

Nguyen, D. Q., Billingsley, R., Du, L., & Johnson, M. (2015). Improving topic models with latent feature word representations. Transactions of the Association for Computational Linguistics, 3, 299–313.

Porter, A., & Chubin, D. (1985). An indicator of cross-disciplinary research. SCIENTOMETRICS, 8(3–4), 161–176. Amsterdam: Elsevier Science Bv.

Prabowo, R., & Thelwall, M. (2006). A comparison of feature selection methods for an evolving RSS feed corpus. Information Processing & Management, Special Issue on Informetrics, 42(6), 1491–1512.

Price, D. (1963). Little science, big science. Columbia University Press.

Qaiser, S., & Ali, R. (2018). Text mining: Use of TF-IDF to examine the relevance of words to documents. International Journal of Computer Applications, 181.

Rettenmeier, L. (2020, July 23). Word embeddings: Stability and semantic change. arXiv. Retrieved January 13, 2025, from http://arxiv.org/abs/2007.16006

Salton, G., Allan, J., & Singhal, A. (1996). Automatic text decomposition and structuring. Information Processing & Management, 32(2), 127–138.

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379–423. Presented at the The Bell System Technical Journal.

Tien, N. H., Le, N. M., Tomohiro, Y., & Tatsuya, I. (2019). Sentence modeling via multiple word embeddings and multi-level comparison for semantic textual similarity. Information Processing & Management, 56(6), 102090.

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. Information Processing & Management, Patent Processing, 43(5), 1216–1247.

Vicente-Gomila, J. M., Artacho-Ramirez, M. A., Ting, M., & Porter, A. L. (2021). Combining tech mining and semantic TRIZ for technology assessment: Dye-sensitized solar cell as a case. Technological Forecasting and Social Change, 169, 120826.

Wang X., Lu L., & Tai W. (2015). Research of a New Algorithm of Words Similarity Based on Information Entropy. Computer Technology and Development, 25(9), 119–122.

Xu, H., Guo, T., Yue, Z., Ru, L., & Fang, S. (2016). Interdisciplinary topics of information science: A study based on the terms interdisciplinarity index series. SCIENTOMETRICS, 106(2), 583–601.

Yan, L. U. O., Shuliang, Z., Xiaochao, L. I., Yuhui, H. a. N., & Yafei, D. (2016). Text keyword extraction method based on word frequency statistics. Journal of Computer Applications, 36(3), 718.

Yang, J., Liu, Y., Zhu, X., Liu, Z., & Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Information Processing & Management, 48(4), 741–754.

Yao, R., Wang, J., Wu, J., Xu, Q., & Zhang. (2023). Research on potential interdisciplinary topic identification method. Library and Information Service, 67(15), 80–93.

Yu, Y., Ju, P., & Shang, M. (2022). Research on the evaluation method of patent keyword extraction algorithm based on information gain and similarity. Library and Information Service, 66(6), 108–117.

Yu Yan & Zhao Naixuan. (2018). Weighted topic model for patent text analysis. Data Analysis and Knowledge Discovery, 2(4), 81–89.

Zhang, L., Sun, B., Chinchilla-Rodríguez, Z., Chen, L., & Huang, Y. (2018). Interdisciplinarity and collaboration: On the relationship between disciplinary diversity in departmental affiliations and reference lists. Scientometrics, 117(1), 271–291.

Zhao, M., Guo, J., & Wu, X. (2024). A large group emergency decision-making approach on HFLTS with public preference data mining. Journal of Global Information Management (JGIM), 32(1), 1–22.

https://doi.org/10.51408/issi2025_028

A Novel Bibliometric Algorithm Unveils the Prevalence and Significance of Gender Match in Research Mentorship Networks

David Campbell[1], Guillaume Roberge[2], Elisabeth Browning[3]

[1]*d.campbell@elsevier.com,* [2]*g.roberge@elsevier.com,* [3]*e.browning@elsevier.com*
Analytical Services, Elsevier (Canada)

Abstract

Gender homophily in research mentorship was investigated across a selection of countries and research fields over two decades by leveraging a novel tree algorithm that uncovers mentor–mentee links from genderized author profiles in Scopus. Despite a decrease in homophily for women relative to men, the overrepresentation of same-gender pairs remained much higher for women than for men in 2022. Only in fields where women were the dominant group was homophily in mentorship more pronounced for men than for women. Next, the contribution, relative to expectations, of same- and opposite-gender mentees to the tree index of their mentors was explored. This new metric quantifies the contribution of mentors to the future bibliometric performance of their mentees. Pairing with same-gender mentors was found to slightly and nearly systematically benefit the future bibliometric performance of women mentees across countries and research fields, regardless of their status as the underrepresented gender. While the positive impact on the performance of women mentees is small, the consistent pattern observed across countries and fields suggests that this is a genuine effect. The robustness of our findings across different contexts suggests that the availability of same-gender mentors is more critical for women than for men, due to women's lower representation in most areas of science. In contrast, the bibliometric performance of men mentees only appeared to benefit from a gender match in mentorship in the few subfields in which men are underrepresented. These results underscore the importance of gender match in research mentorship networks, particularly for women mentees, as well as critical aspects of the dynamics of research mentorship networks.

Introduction

Research mentorship refers to a multifaceted relationship where experienced researchers (mentors) guide less experienced individuals (mentees) in their personal and professional development. Such mentorship, when built on mutual respect and commitment, has been shown to significantly impact the professional trajectories of the next generation of scholars, enhancing their technical (e.g., research design, instrument use, data treatment and analysis) and non-technical skills (e.g., networking, authorship practices, fundraising, mentoring) in various aspects of research. This, in turn, contributes to the success of early career researchers (e.g., graduate students) as they move on to independent research careers (Guston, 1993; National Academies of Sciences-Engineering and Medicine, 2019).

Although research mentoring is often regarded as being within the remit of formal roles (e.g., supervisors), other researchers, such as collaborators, can fulfill informal mentorship roles through the provision of guidance and support (e.g., experience sharing, offering feedback) to less experienced individuals. Having multiple mentors can enrich a mentee's experience, notably by broadening perspectives and networking opportunities (Atenas et al., 2023; Gorela & Biloslavo, 2015).

A growing body of research suggests that same-gender mentor–mentee relationships within various academic settings are far more common than expected under a gender-neutral pairing assumption (Gallen & Wasserman, 2023; Moghe et al., 2021; Morales et al., 2018; Murphy et al., 2022; Schwartz et al., 2022). Furthermore, matching gender in mentor–mentee relationships, especially for women, could more effectively equip mentees for an academic career, yielding greater benefits than when mentored by someone of the opposite gender (Morales et al., 2018; Schwartz et al., 2022).

However, these studies were limited in geographic and/or disciplinary scope and were reliant on local surveys or databases (e.g., online mentoring platforms or databases indexing theses (COBISS)) to uncover and gather information on mentor–mentee relationships. Perhaps the most extensive analysis was performed by Schwartz and colleagues (2022) using data from the Academic Family Tree (https://www.academictree.org), though this was limited to the life sciences and had a predominantly US focus.

This study aims to confirm and assess the generalizability of existing evidence on the extent of gender homophily in research mentorship networks, as well as the potential benefits, in terms of bibliometric performance, of a gender match in mentorship. This is achieved by leveraging a novel tree algorithm and tree index (Roberge et al., 2024). By uncovering both formal and informal mentorship relationships from genderized author profiles in Scopus, the tree algorithm enables a large-scale examination of the dynamics of research mentorship networks over the past two decades, as well as across several countries and all fields of science to address the following questions:

- **Q1:** Are research mentorship networks gender homophilic?
- **Q2:** Is gender homophily in research mentorship networks more pronounced for the underrepresented gender?
- **Q3:** Is gender homophily among women researchers (usually the underrepresented gender) negatively correlated with their representation in research?
- **Q4:** Has gender homophily among women researchers (usually the underrepresented gender) declined as their representation in research increased over the past 20 years?

By capturing the average cumulative performance of a mentor's mentees as independent researchers later in their careers, the tree index enables assessing the benefits of same-gender mentorships on the subsequent bibliometric performance of mentees to address the following questions:

- **Q5:** Is gender match in mentorship beneficial to the bibliometric performance of mentees, as evidenced by the publications they produce independently of their mentors later in their career?
- **Q6:** Are performance gains from same-gender mentorships more pronounced for the underrepresented gender?

The paper concludes with a discussion of the implications of the study findings in light of the existing literature on gender homophily in research.

Methodology

Data source

The Scopus database produced by Elsevier includes abstracts and citation information from more than 90 million records covering all fields of science and technology, including the social sciences, arts and humanities. The September 1st 2023 snapshot of Scopus (Elsevier) was used to retrieve all necessary metadata on peer-reviewed scientific publications, mainly articles, conference papers, reviews, and short surveys published in book series, conference proceedings, or journals having valid ISSNs. Hereafter, these records are referred to as publications or papers.

Disambiguated author profiles

The novel tree algorithm and tree index used in this study rely on Scopus author IDs (AUIDs). The disambiguated AUIDs offer a clean view of a researcher's full publication history as indexed in Scopus. Using a large-scale "gold set" of over 10,000 manually cleaned author profiles, Campbell & Struck (2019) have shown that Scopus AUIDs enable robust conclusions in an evaluative context of groups of at least 500 researchers. They estimated the recall and precision of the AUIDs at, respectively, 98% and 96.9%.

Gender inference

The scale of the analysis pursued in this study, spanning several countries and scientific disciplines, required getting gender data for millions of Scopus authors. Collecting self-declared gender data at such a scale has never been undertaken and would not have been feasible within the time and budgetary constraints of the study. Therefore, the study team opted to infer binary gender, recognizing the limitation that this approach does not account for all gender groups. Additionally, had data been available for non-binary gender groups, including them could have risked identifying specific individuals as they represent a very small proportion of the population.
The Namsor API was used to infer the binary gender of all authors (covering mentors and mentees) in Scopus. Authors were classified as a man or a woman if the probability of being a man or a woman exceeded 85%. Pinheiro, Durning, & Campbell (2022) demonstrated that results were robust to changes in this gender assignation rule in a multivariate analysis of the relationship between gender and interdisciplinarity. Additionally, this is a well-established method that has been used in several rounds of *She Figures* by the European Commission (2021, 2024).

Tree algorithm

In this study, a tree algorithm was used to identify both formal and informal mentorship ties through an examination of key co-authorship patterns between senior and junior researchers, as summarized below.

The tree algorithm identifies mentor–mentee relationships from the publication history of Scopus AUIDs as follows (refer to Roberge et al. (2024) for more information):

- Senior researchers, defined as those with at least 10 years of publishing experience and at least one paper in the year under investigation, are identified.
- Researchers in Scopus are assigned as a potential mentor the senior researcher with whom they have co-published the most within their first five years as authors in Scopus. A minimum of 2 co-publications is required for mentor assignment, and ties can result in multiple mentors.
- Only mentees who later (after the first five years) published at least two papers independently from their potential mentor(s) are retained in the mentor's tree, ensuring it only includes mentees who had some success in a subsequent publishing career.

Roberge et al. (2024) utilized the E-Theses Online Service (EThOS)[1] of the National Library of the UK to validate the "mentoring" character of the identified mentor–mentee links using metadata from over 100,000 theses (mostly PhDs) awarded between 1980 and 2022. Excluding student–supervisor pairs from EThOS where the tree algorithm could not assign any mentor to the students matched to Scopus, the share of students linked to their correct supervisor(s) was 77% (ranging from 67% to 83% across disciplines, with lower accuracies observed in the Social Sciences and Humanities (SSH)). While validation using EThOS data showed that the tree algorithm frequently captures formal mentorship ties in the form of supervisor–student pairs, it is worth noting that it also captures other senior–junior interactions. In this paper, such collaborations between senior and junior researchers are assumed to have been accompanied by informal peer mentoring (e.g., experience sharing) known to occur in research collaboration networks. Future research could test this assumption by asking a sample of senior researchers to review the list of their non-student mentees as identified by the tree algorithm.

Using EThOS data, Roberge et al. (2024) also showed that in some disciplines, mainly in the SSH, supervisor–student pairs are not captured as frequently, likely due to coverage issues in Scopus (the bibliographic database used in this study). As a result, this study's findings for the SSH may not be as robust as for other disciplines.

Time series

Running the tree algorithm requires substantial computational resources. To analyze trends, data have therefore been generated for a limited set of years over the last 20 years, specifically in 2002, 2006, 2010, 2014, 2018, and 2022. In any of these years, the tree algorithm accounted for the full publication history of a mentor, and its mentees, up to that year (inclusive).

[1] https://bl.iro.bl.uk/concern/datasets/308c54ce-31b1-4cb1-b257-7b288a3c7926?locale=en

Country and field coverage

For each of the above years, the mentors have been limited to authors who qualified as senior and actively published at least one paper in the corresponding year. Mentors were uniquely assigned to the country and subfield in which they published most of their publications up to that year.[2] Although researchers might have visited more than one country, the homophily signature linked to a given mentor should mostly reflect the situation in the main country of affiliation. When aggregating across a country's mentors, the extent to which they contributed to homophily is expected to converge to the country's main pattern minimizing noise from the secondary countries of mentors.

To test the robustness of the study findings, as well as to assess their generalizability, the analyses were repeated for 38 countries (EU27 members plus Argentina, Australia, Brazil, Canada, Egypt, India, Japan, Mexico, South Africa, United Kingdom, and the United States)[3] overall in Scopus and by main field of research based on Science-Metrix's classification.[4]

Some analyses were repeated for a small group of subfields with a majority of woman mentors (i.e., Gender Studies, Nursing, Nutrition & Dietetics, Developmental & Child Psychology, Public Health and Social Work). This was done to assess whether gender differences in mentorship homophily and associated performance gains differ from the dominant patterns where male researchers are in the majority.

Homophily indicator

To test hypotheses 1 to 4, the study examined, across the selected fields, countries, and years, the extent to which the frequency of same-gender links departs from expectation under a gender-neutral (random) pairing assumption by gender of the mentors.

For each mentor in a given year, all prior mentees up to that year are considered regardless of originating country(ies), as students may come from abroad. Therefore, the expected share of women and men mentees in a given field, country and year for each mentor, regardless of gender, is based on the pooled set of mentees of a country's mentors, including those from abroad, in the corresponding field and year. As an example, if 44% of the pooled mentees of women and men mentors in Canada are women, one would expect 44% of the mentees of women mentors to be women if the assortment of mentors and mentees was gender neutral. A share of woman−woman links above 44% would denote a homophilic network from the perspective of women mentors. In Canada, 60% of the mentees of women mentors are women leading to a positive deviation of 36% relative to expectation (homophily for women mentors = (0.60/0.44) - 1 = 0.36). The same approach was applied in

[2] In the rare case of ties, mentors were randomly assigned to one of the tied countries and were assigned to all the tied subfields. We will further test the impact on the study conclusions by either excluding mentors or assigning them to multiple countries in the case of ties.

[3] China was not included due to issues in assigning gender to authors.

[4] https://www.science-metrix.com/classification/

exploring the extent of deviation from expectation for man–man links (homophily for men mentors = (0.61/0.56) - 1 = 0.09).

Tree index

To address questions 5 and 6, the contribution of same- and opposite-gender mentees to the tree index of their mentors was investigated across the selected fields, countries, and years.

The tree index is a new subfield- and year-normalized metric designed to quantify the impact of mentors on the bibliometric performance of the next generation of scientists. For each mentor, this composite indicator accounts for the cumulative volume and impact of his or her mentees' publications, as well as the size of their co-authorship network, as they go on to independent careers (refer to Roberge et al. (2024) for more information). As such, each mentee contributes a certain share to the mentor's tree index.

If mentees of women mentors consist, on average, of 61.21% women in Canada, and the average share of these mentors' tree index that is attributable to women mentees equals 60.76%, the tree index departure from expectation for woman–woman (mentee–mentor) pairings would equal -0.7% (0.608/0.612 - 1). Applying the same approach to other pairing types produces the following deviations for Canada in Scopus (2022): -0.7% woman–woman, -2.8% woman–man, 1.7% man–man, and 1.1% man–woman.

As observed for Canada, woman mentees are likely to contribute less than expected to their mentors' tree index regardless of the gender of their mentors because of gender inequalities in research (European Commission, 2021). The opposite applies to men mentees. Nevertheless, by comparing the average departure from expectation in the contribution of mentees to the tree index of their women and men mentors, the study enables an assessment of whether same-gender mentorships are associated with performance gains for both women and men mentees.

Note that expectation assumes all else is equal even if not the case (e.g., publishing age of a mentor's mentees). This is later accounted for in interpreting the results.

Results

Descriptive statistics on the study data set

Among the selected countries (all fields combined), the share of mentors and mentees with unknown gender was small, with no major implication for the study findings (Figure 1, left). As expected, the share of women mentors and mentees has been steadily increasing over the past two decades (Figure 1, right). The share of women among the approximately 2.4 million mentees was unsurprisingly higher than their share among the approximately 0.6 million mentors in 2022. This is due to the well-known leaky pipeline and glass ceiling phenomena in academia, whereby a smaller share of women researchers reaches senior levels (European Commission, 2021).

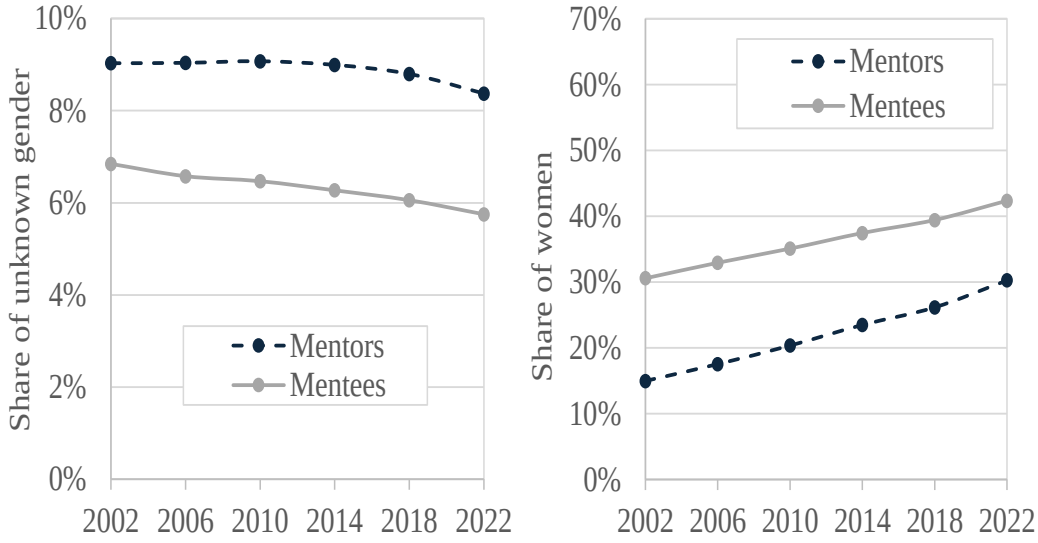


Figure 1. Trends in the average share of unknown gender (left) and women (right) among mentors and mentees of the selected countries, 2002–2022.

Note: The share of women is calculated excluding unknowns.

Questions 1 to 4

The percent deviation in the share of same-gender mentees relative to neutral expectation (random assignment without regard to gender) is depicted by country and gender of the mentors in Figure 2, for all fields combined in 2022. Some key patterns emerge. First, over the course of mentors' careers, the share of same-gender mentees is systematically higher than expected for both men and women mentors active in 2022. Second, and more interestingly, the degree of homophily is systematically more pronounced for women, usually the underrepresented gender, than for men mentors. On average across the selected countries, woman–woman links are overrepresented by 43% relative to expectation, versus 11% for man–man links. There is also a moderate negative correlation (-0.53) between the share of women mentees and the homophily of women mentors across countries (data not shown). Thus, with greater representation of women in research, one might expect a decrease in homophily within research mentorship networks, a pattern that is indeed observed in the study's results.

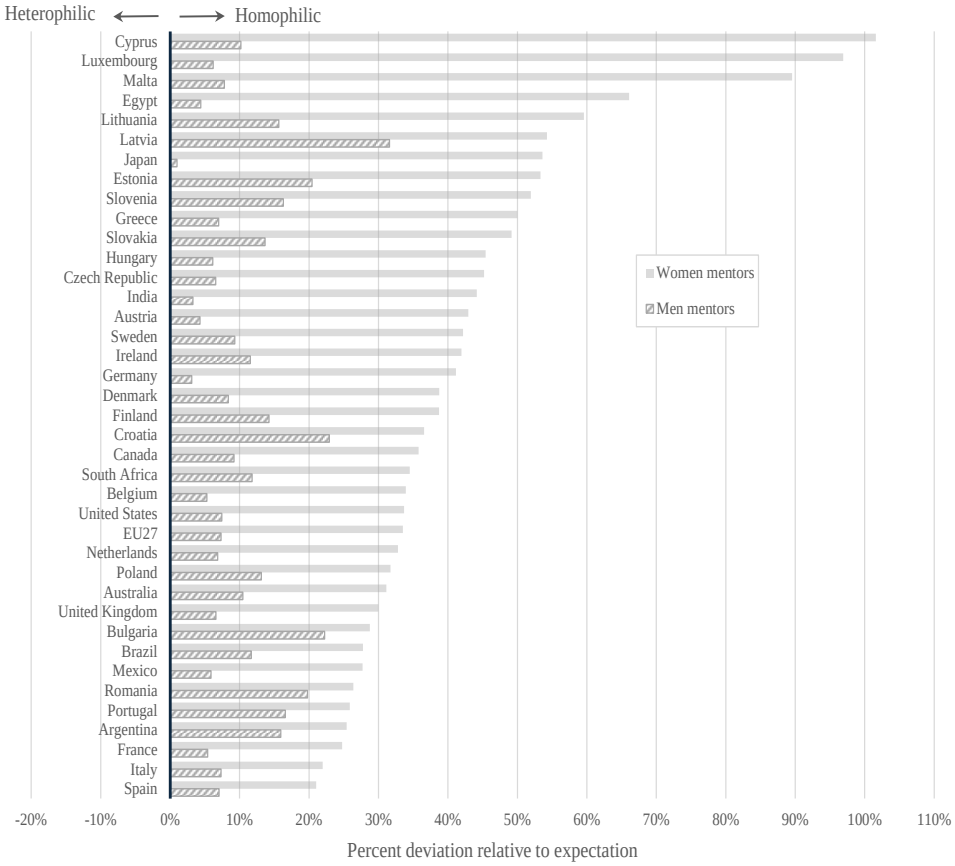


Figure 2. Percent deviation in the share of same-gender mentees relative to expectation in the overall research mentorship network, by country and gender of the mentor (2022).

Across the selected countries (all fields combined), the extent of homophily systematically decreased for women (average CAGR of -4.0%) as their representation increased over the past two decades (Table 1). This pattern was matched by an opposite trend of the same magnitude for men (average CAGR of +3.8%). The increase for men was also systematic across countries.

Table 1. Trends in percent deviation in the share of same-gender mentees relative to expectation in the overall research mentorship network, by country and gender of the mentor, 2002–2022.

	Women mentors				Men mentors			
Country	**2002**	**2022¥**	**CAGR**	**Trend***	**2002**	**2022**	**CAGR**	**Trend***
Egypt	185%	66%	-5.0%		3.1%	4.4%	1.7%	
Japan	59%	54%	-0.5%		0.4%	1.0%	5.0%	
Slovenia	121%	52%	-4.1%		8.1%	16.3%	3.6%	
Greece	157%	50%	-5.6%		3.9%	7.0%	2.9%	
Slovakia	113%	49%	-4.1%		8.3%	13.7%	2.5%	
Hungary	93%	45%	-3.5%		3.6%	6.2%	2.6%	
Czech Republic	107%	45%	-4.2%		5.8%	6.6%	0.7%	
India	80%	44%	-2.9%		2.3%	3.3%	1.7%	
Austria	123%	43%	-5.1%		1.2%	4.3%	6.6%	
Sweden	84%	42%	-3.4%		3.0%	9.3%	5.9%	
Ireland	138%	42%	-5.8%		3.3%	11.6%	6.5%	
Germany	99%	41%	-4.3%		0.8%	3.1%	7.2%	
Denmark	77%	39%	-3.4%		2.9%	8.4%	5.5%	
Finland	68%	39%	-2.8%		7.1%	14.2%	3.6%	
Croatia	82%	37%	-3.9%		13.3%	22.9%	2.8%	
Canada	86%	36%	-4.3%		3.2%	9.2%	5.4%	
South Africa	142%	34%	-6.8%		4.0%	11.8%	5.5%	
Belgium	87%	34%	-4.6%		2.1%	5.3%	4.8%	
United States	66%	34%	-3.3%		2.8%	7.4%	5.0%	
EU27	71%	33%	-3.7%		3.5%	7.3%	3.7%	
Netherlands	68%	33%	-3.6%		1.1%	6.9%	9.6%	
Poland	55%	32%	-2.7%		9.9%	13.1%	1.4%	
Australia	75%	31%	-4.3%		3.6%	10.5%	5.4%	
United Kingdom	62%	30%	-3.6%		2.3%	6.6%	5.4%	
Bulgaria	57%	29%	-3.4%		13.8%	22.3%	2.4%	
Brazil	70%	28%	-4.5%		7.8%	11.7%	2.0%	
Mexico	94%	28%	-5.9%		5.2%	5.9%	0.7%	
Romania	63%	26%	-4.2%		10.5%	19.8%	3.2%	
Portugal	73%	26%	-5.0%		10.6%	16.6%	2.3%	
Argentina	57%	25%	-4.0%		14.5%	15.9%	0.5%	
France	42%	25%	-2.6%		3.4%	5.4%	2.3%	
Italy	36%	22%	-2.4%		4.1%	7.3%	2.9%	
Spain	45%	21%	-3.7%		3.9%	7.0%	3.0%	

Note: ¥Countries are sorted based on the extent of homophily for women mentors. *The trends are on a common scale to show that the absolute magnitude of change in homophily is smaller for men than women mentors despite the average magnitude of their CAGR being very similar, yet in opposite direction. Some countries are excluded due to too few observations.

Finally, when disaggregating the data presented in Figure 2 by field of science (data not shown), 99% of observations confirm the tendency towards same-gender pairings for both women and men mentors. The pattern is also more pronounced for women than men mentors 86% of the time. Interestingly, among the few subfields in which most mentors are women (i.e., Gender Studies, Nursing, Nutrition & Dietetics, Developmental & Child Psychology, Public Health and Social Work) and for which there is enough data (selected countries pooled) to analyse gender homophily in mentorship, the pattern of greater same-gender pairing was inverted. In these subfields (grouped), in 2022, a tendency to same-gender pairing is more pronounced for men (+25% deviation) than for women (+7%) (Figure 3).

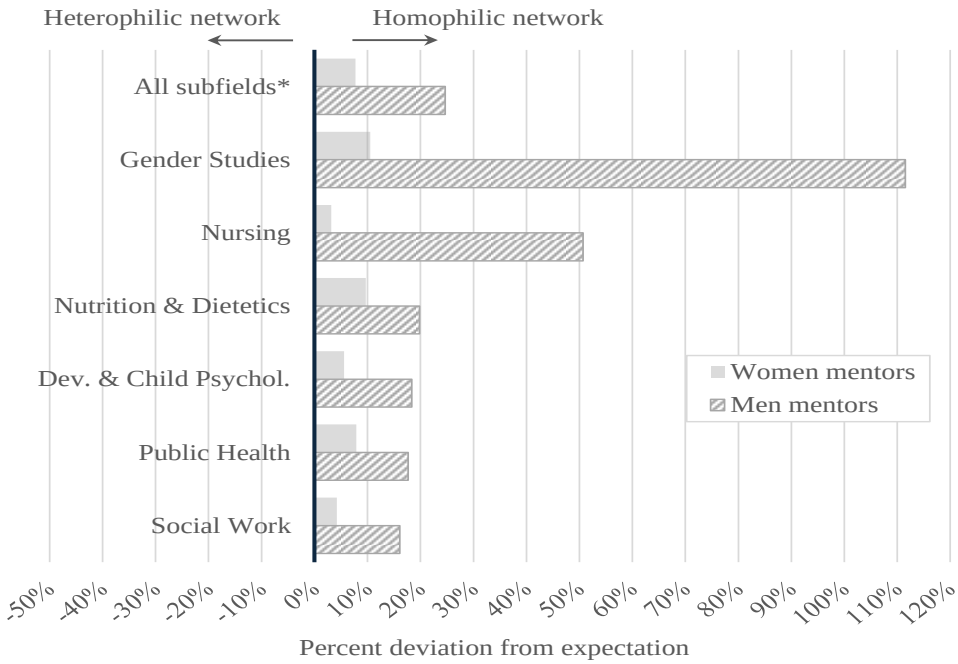


Figure 3. Percent deviation from expectation in the share of same-gender mentees in the overall research mentorship network, by subfield in which most mentors are women and gender of the mentor (2022).

Note: *All subfields are all those with a majority of women mentors grouped.

Questions 5 & 6

Accounting for those mentors who were actively publishing in each of the selected countries in 2022 (all fields combined), Figure 4 presents the extent to which the contribution of women and men mentees to the tree index of women and men mentors deviated from expectations. The key gender differences highlighted below and in Figure 4 do not appear to be due to gender differences in the average age of a mentor's mentees (data not shown):

- As anticipated, due to long-standing inequalities in research, women mentees contributed less than expected to the tree index of their mentors regardless of their mentors' gender. The opposite was observed for men mentees. Although

these departures from expectation are of a small magnitude, they are nearly systematic with only a handful of exceptions (i.e., Latvia, Estonia and Luxembourg).

- The underperformance of women mentees relative to expectation when paired with mentors of the same gender is, on average, roughly less than half that observed when they are paired with mentors of the opposite gender. This pattern is also systematic across all selected countries in all fields combined (except for Luxembourg and Lithuania) and was repeated in 75% of all country−field combinations (data not shown).
- A similar result is not consistently observed for men mentees for whom being paired with a mentor of the same gender only equated to performance gains for about half of the countries examined (18 out of 38 countries).

Despite the small magnitude of observed deviations from expectations, their consistency across countries and fields in 2022 suggests that same-gender mentors for junior women researchers offer some benefits, even if only slightly. This consistency warrants a deeper investigation into the root causes of this finding to uncover ways in which men mentors could better support women mentees.

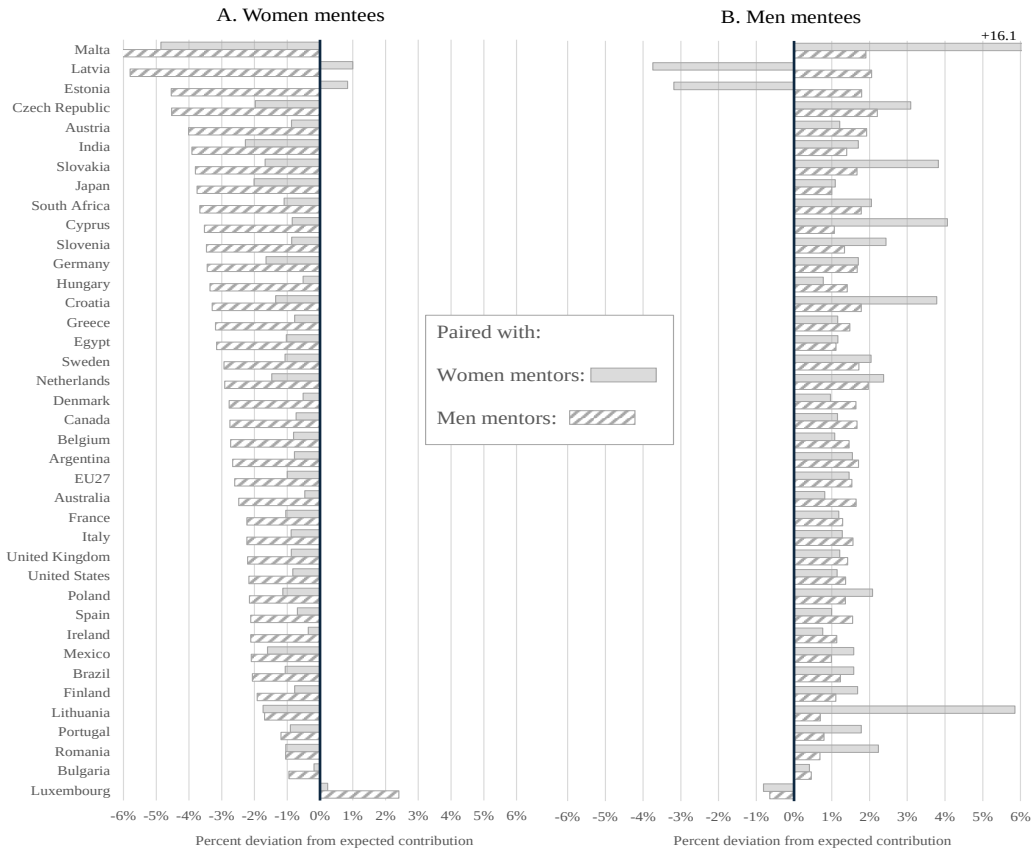


Figure 4. Percent deviation from expected contribution of women/men mentees to the tree index of women/men mentors in the overall research mentorship network, by country (2022).

Although the main patterns in network homophily were inverted for subfields with a majority of women mentors (see Figure 3), the patterns were not fully inverted in terms of the contribution of mentees to the tree index of mentors (Figure 5). The closest match to perfect inversion was observed in Gender Studies, where men mentees underperformed and women mentees overperformed regardless of the gender of their mentors. In that subfield, the negative deviation for men mentees was also less pronounced when paired with men than women mentors. However, in all other subfields, women mentees still underperformed relative to expectation and benefited from being paired with women mentors, although the effect sizes were still very small. Interestingly, in these subfields, even though men mentees systematically overperformed relative to expectation, they also appeared to systematically benefit from being paired with same-gender mentors, a pattern that was not observed in areas with a majority of men mentors.

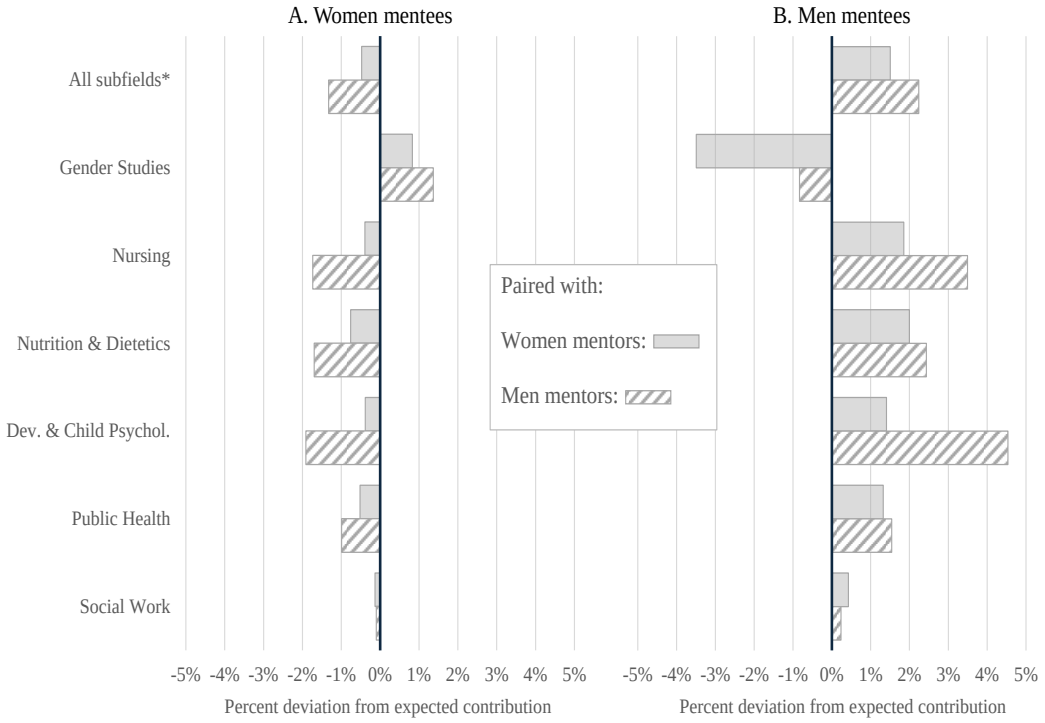


Figure 5. Percent deviation from expected contribution of women/men mentees to the tree index of women/men mentors in the overall research mentorship network, by subfields with a majority of women mentors (2022).

Note: *All subfields are all those with a majority of women mentors.

Discussion

Research mentorship networks were shown to be gender homophilic across nearly all country–field combinations for both women and men, generalizing findings from smaller-scale studies (Gallen & Wasserman, 2023; Moghe et al., 2021; Morales et al., 2018; Murphy et al., 2022; Schwartz et al., 2022). This suggests that mentors,

regardless of their gender, likely play an important role in attracting same-gender mentees to research and mentoring them later on.

Results also confirm that gender homophily in mentorship is usually more pronounced for the underrepresented gender. Homophily was more pronounced among women, typically the underrepresented gender, than among men researchers in 86% of the country–field combinations examined. This pattern was also observed in research collaboration networks (Hajibabaei et al., 2022; Wang et al., 2023), although not systematically (Kwiek & Roszka, 2021). Additionally, this pattern was inverted (i.e., greater homophily among men than women) in the few subfields where most mentors are women and for which there was enough data to analyse gender homophily in mentorship.

The extent of gender homophily among women was also negatively correlated with their representation in research. Women were more likely to pair with other women in country–field combinations where they were more underrepresented. Although the correlation was moderate, additional observations suggest that the more pronounced homophily among women could be driven by their status as the underrepresented group. Over the past two decades, the extent of homophily among women decreased with their increased representation in research, and the extent of homophily was found to be greater among men in the few areas of science where they are underrepresented (i.e., Gender Studies, Nursing, Nutrition & Dietetics, Developmental & Child Psychology, Public Health and Social Work).

The study's results therefore suggests that the availability of same-gender mentors is more critical for women than for men, due to women's lower representation in most areas of science. The importance of gender match in mentorship is further exemplified by the nearly systematic, albeit small, positive impact of same gender mentors on the later research performance of women mentees. Although the bibliometric performance of women mentees appeared to benefit from same-gender mentors regardless of their status as the underrepresented gender, men mentees only appeared to benefit from same-gender mentors in the few areas in which they were underrepresented. This is consistent with prior studies reporting benefits, especially for women, of matching gender in mentor−mentee relationships (Moghe et al., 2021; Morales et al., 2018; Schwartz et al., 2022).

It seems possible that members of underrepresented groups might seek out others with similar experiences and concerns about potential issues with the majority group. However, as diversity increases, these obstacles might diminish. Research by Bai, Ramos, & Fiske (2020) has shown that "as diversity increases, people paradoxically perceive social groups as more similar," possibly leading to fewer stereotypes.

In turn, as gender becomes more equally represented in science, the extent of homophily would be expected to decrease. This underscores the importance of retaining senior women researchers, not only for their direct contributions to science but also for their indirect mentorship contributions via their mentees.

Given the study's limitations as detailed throughout the methods section (e.g., use of inferred binary instead of self-declared non-binary gender, need to further validate the "mentoring" connection of uncovered informal peer mentoring ties), further work relying on a mix of qualitative and quantitative approaches would be warranted to

confirm our findings on homophily and its implications for mentees. However, the robustness of our findings is supported by the parallel results obtained across countries and fields, reinforcing the validity of our conclusions.

These findings should guide policymakers in initiatives aimed at encouraging the greater participation of women in science. For instance, interventions to increase the participation of women in science, especially in countries or fields, such as science, technology, engineering, or math where they are heavily underrepresented, appear highly relevant considering the study's results. Interventions directed towards increasing the retention rates of women as they advance through academic careers may be particularly effective. In these cases, concomitant interventions to strengthen the mentoring skills of men towards women may also be warranted, and further research could help uncover the main levers for intervention.

Altogether, increasing gender diversity in research, and in research teams, should be the ultimate target as several recent studies underscore the unique value of mixed-gender teams in fostering novel, disruptive, and influential scientific discoveries (Hajibabaei et al., 2022; Yang et al., 2022; Zhang et al., 2024).

Acknowledgements

The authors would like to thank senior management at Elsevier, in particular Nicolien van der Linden, M'hamed el Aisati and Olivier Dumon, for their interest and support in leveraging the novel tree algorithm and tree index to address key questions of high relevance to promoting an inclusive and diverse research culture at all stages of an academic career. This research was funded through mandated work conducted for the European Commission Directorate General for Research and Innovation.

Author contributions

Conceptualization: David Campbell
Methodology: David Campbell, Guillaume Roberge
Data curation: David Campbell, Guillaume Roberge
Formal Analysis: David Campbell
Writing – original draft: David Campbell
Writing – review & editing: Elisabeth Browning, David Campbell, Guillaume Roberge

Competing interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the working process

The authors used Scopus AI to assist the literature search. During preparation of the manuscript, the authors used Microsoft Copilot, built upon OpenAI's GPT-4, to improve readability and language of certain sentences. The authors carefully reviewed and edited the proposed changes as needed and take full responsibility for the content of the publication.

References

Atenas, J., Nerantzi, C., & Bussu, A. (2023). A Conceptual Approach to Transform and Enhance Academic Mentorship: Through Open Educational Practices. *Open Praxis*, *15*(4), 271–287. https://doi.org/10.55982/OPENPRAXIS.15.4.595

Bai, X., Ramos, M. R., & Fiske, S. T. (2020). As diversity increases, people paradoxically perceive social groups as more similar. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(23), 12741–12749. https://doi.org/10.1073/PNAS.2000333117/SUPPL_FILE/PNAS.2000333117.SAPP.PDF

Campbell, D., & Struck, B. (2019). Reliability of Scopus author identifiers (AUIDs) for research evaluation purposes at different scales. *17th International Conference of the International Society for Scientometrics and Informetrics (ISSI), 2–5 September 2019, Proceedings Vol. II*, 1276–1287. http://issi-society.org/publications/issi-conference-proceedings/proceedings-of-issi-2019/

European Commission. (2021). *She Figures 2021*. https://research-and-innovation.ec.europa.eu/knowledge-publications-tools-and-data/publications/all-publications/she-figures-2021_en

European Commission. (2024). *She Figures 2024*. Upcoming

Gallen, Y., & Wasserman, M. (2023). Does information affect homophily? *Journal of Public Economics*, *222*, 104876. https://doi.org/10.1016/J.JPUBECO.2023.104876

Gorela, K., & Biloslavo, R. (2015). Relationship between Senior and Junior Researcher: Challenges and Opportunities for Knowledge Creating and Sharing. In P. Diviacco, P. Fox, C. Pshenichmy, & A. Leadbetter (Eds.), *Collaborative Knowledge in Scientific Research Networks* (pp. 90–125). IGI Global. https://doi.org/10.4018/978-1-4666-6567-5.CH006

Guston, D. H. (1993). Mentorship and the Research Training Experience. In National Academy of Sciences, National Academy of Engineering and Institute of Medicine (Ed.), *Responsible Science: Ensuring the Integrity of the Research Process: Volume II.* (pp. 50–65). National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK236193/

Hajibabaei, A., Schiffauerova, A., & Ebadi, A. (2022). Gender-specific patterns in the artificial intelligence scientific ecosystem. *Journal of Informetrics*, *16*(2), 101275. https://doi.org/10.1016/j.joi.2022.101275

Kwiek, M., & Roszka, W. (2021). Gender-based homophily in research: A large-scale study of man-woman collaboration. *Journal of Informetrics*, *15*(3), 101171. https://doi.org/10.1016/J.JOI.2021.101171

Moghe, S., Baumgart, K., Shaffer, J. J., & Carlson, K. A. (2021). Female mentors positively contribute to undergraduate STEM research experiences. *Volume 16, Issue 12 December*, *16*(12 December). https://doi.org/10.1371/journal.pone.0260646

Morales, D. X., Grineski, S. E., & Collins, T. W. (2018). Effects of gender concordance in mentoring relationships on summer research experience outcomes for undergraduate students. *Volume 102, Issue 5, Pages 1029 - 1050*, *102*(5), 1029–1050. https://doi.org/10.1002/sce.21455

Murphy, M., Record, H., Callander, J. K., Dohan, D., & Grandis, J. R. (2022). Mentoring Relationships and Gender Inequities in Academic Medicine: Findings from a Multi-Institutional Qualitative Study. *Academic Medicine*, *97*(1), 136–142. https://doi.org/10.1097/ACM.0000000000004388

National Academies of Sciences-Engineering and Medicine; Policy and Global Affairs; Board on Higher Education and Workforce; Committee on Effective Mentoring in STEM. (2019). The Science of Mentoring Relationships: What Is Mentorship? In M. L. Dahlberg & A. Byars-Winston (Eds.), *The Science of Effective Mentorship in STEMM*

(pp. 33–50). National Academies Press (US). https://doi.org/10.17226/25568
Pinheiro, H., Durning, M., & Campbell, D. (2022). Do women undertake interdisciplinary research more than men, and do self-citations bias observed differences? *Quantitative Science Studies, 3*(2), 363–392. https://doi.org/10.1162/qss_a_00191
Roberge, G., Campbell, D., Browning, E., Dong, D., Khayat, P., El Aisati, M., & Dumon, O. (2024). Tree Index: a new widescale indicator on contribution to mentorship. *Proceedings of the 2024 STI Conference, Berlin, 18-20 September*. https://zenodo.org/records/13987541
Schwartz, L. P., Liénard, J. F., & David, S. V. (2022). Impact of gender on the formation and outcome of formal mentoring relationships in the life sciences. *Volume 20, Issue 9, 20*(9). https://doi.org/10.1371/journal.pbio.3001771
Wang, Y. S., Lee, C. J., West, J. D., Bergstrom, C. T., & Erosheva, E. A. (2023). Gender-based homophily in collaborations across a heterogeneous scholarly landscape. *PLoS ONE, 18*(4). https://doi.org/10.1371/journal.pone.0283106
Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences of the United States of America*, 119(36), e2200841119. https://doi.org/10.1073/PNAS.2200841119/SUPPL_FILE/PNAS.2200841119.SAPP.PDF
Zhang, M. Z., Wang, T. R., Lyu, P. H., Chen, Q. M., Li, Z. X., & Ngai, E. W. T. (2024). Impact of gender composition of academic teams on disruptive output. *Journal of Informetrics*, 18(2), 101520. https://doi.org/10.1016/j.joi.2024.101520

https://doi.org/10.51408/issi2025_057

A Research Entities Disambiguation Methodology Tested on Brazilian Researchers Database

Alysson Fernandes Mazoni[1], Estevão Fernandes Macedo[2], Luís Fabiano Farias Borges[3], Esteban Fernandez Tuesta[4]

[1]*afmazoni@unicamp.br*
University of Campinas, Department of Science and Technology Policy, Institute of Geosciences, R. Carlos Gomes, 250, Cidade Universitária, Campinas, São Paulo (Brazil)

[2]*estevao.macedo@usp.br*
University of São Paulo, Interdisciplinary Group of Modeling Complex Systems, School of Arts, Sciences and Humanities, Rua Arlindo Béttio, 1000, Ermelino Matarazzo, São Paulo, São Paulo (Brazil)

[3]*luis.borges@capes.gov.br*
University of Campinas, Department of Science and Technology Policy, Institute of Geosciences, R. Carlos Gomes, 250, Cidade Universitária, Campinas, São Paulo (Brazil)
Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Setor Bancário Norte, Quadra 2, Bloco L, Lote 6, Brasília (Brazil)

[4]*tuesta@usp.br*
University of São Paulo, Interdisciplinary Group of Modeling Complex Systems, School of Arts, Sciences and Humanities, Rua Arlindo Béttio, 1000, Ermelino Matarazzo, São Paulo, São Paulo (Brazil)

Abstract

This paper proposes a methodology for disambiguating research entities in databases, with a focus on matching authors, institutions, and publications across various systems. The study examines the OpenAlex and Lattes databases (Brazil's national researcher registry), aiming to enhance the quality and coverage of both databases. Persistent identifiers, such as DOIs and ORCIDs, are utilized to link entities, while co-authorship and affiliation data assist in the matching process. The Levenshtein distance metric is employed to compare names and titles for accuracy. The proposed method is straightforward to implement in tabular databases, making it an effective solution for research information systems. By improving the linkage of authors and publications, this methodology enhances bibliometric research and data curation on platforms like Lattes and OpenAlex. The results illustrate the potential of integrating local and comprehensive databases to address issues of ambiguous names and incomplete metadata.

Introduction

There are several research entities disambiguation systems and algorithms (Ferreira, Gonçalves, & Laender, 2012; Levin, Krawczyk, Bethard, & Jurafsky, 2012; Sanyal, Bhowmick, & Das, 2021; Xu, Shen, Li, & Fu, 2018). They are used usually inside research information systems in order to allow bibliographic studies. In commercial databases such as Scopus (Boyle & Sherman, 2006) authors are required to fill their data and a persistent identifier is created and maintained that way. That should account for authors and institutions disambiguation. However, the references cited are not always part of the input and tend to not have their metadata correctly

disambiguated. Curatorship of the data is instrumental to their use and much of this work is hidden as a commercial product inside vendors' platforms (Mongeon & Paul-Hus, 2016).

The recent attempts at comprehensive research information databases such as DataCite, OpenAlex, OpenAIRE try to overcome this using several advanced matching algorithms, many of them based on machine learning (Kim & Kim, 2018; Qian, Hu, Cui, Zheng, & Nie, 2011; Rehs, 2021). There are weak spots in these approaches mainly because of coverage that produces incorrect identification (Rehs, 2021).

It is highly likely, as shown in this work, that a projection of the content of these databases on local scientific databases can overcome these problems in regards to ambiguous names, lack of persistent identifiers, among others. In this use case, we used the Lattes database (Mena-Chalco & Junior, 2009) (the Brazilian registry of researchers) as a base to cross their production to OpenAlex. A nearly full matching of researchers, institutions and publications would allow bibliometric research across the Lattes database, that is not fully linked currently and would also correct and improve on the OpenAlex database for it would increase coverage and metadata quality.

In this work, a methodology to match research entities is proposed to disambiguate them inside research information databases. The methodology uses persistent identifiers as clues and their entities links, such as co-authorships and affiliations, to propagate these clues. The clues are then combined using distance metrics for names and titles. The methodology is tested matching research entities from the Brazilian registry of researchers (Curriculum Lattes) against similar entities in the OpenAlex database. The results indicate high precision and ease of implementation and use for tabular databases, which is a common ground for research information systems.

Theoretical Background

Research information systems

A research information system is a sort of database that manages information about scientific activities and production. Institutional databases that keep information on their research such as thesis, monographs, books and articles are examples of such systems (de Castro & Puuska, 2023). Granting agencies usually maintain registries of their supported projects and sometimes try to maintain registries for their research products (Alshamaila et al., 2024). Namely, such systems allow for bibliometric and scientometric research on the cosmos they cover.

There are also databases that purport to be comprehensive about science, in the sense that they aim to cover all knowledge production indexed according to some criteria, such as Scopus, Web of Science, OpenAlex, Dimensions (Turgel & Chernova, 2024).

Such systems are important for the maintaining institutions to be aware of their own function and priorities. In that direction, also government agencies that evaluate scientific research are always in great need of such information, in many cases, keeping such databases as public policy: such is the case of the Lattes platform

(Digiampietri et al., 2012) and other national efforts, CVUY (Simón, Fontáns, & Aguirre-Ligüera, 2013).

Persistent identifiers

The information about scientific activity is spread around several entities, such as authors, journals, institutions, publications, among others. Such entities are usually mapped to data entities in databases in order to create a data model that would allow translating questions about scientific activity into queries or filterings on such databases. Some data models about research are very complex and mature, aiming at traceability of research entities such as the OpenAIRE graph (Vichos et al., 2022) linked to the OpenAIRE database (Manghi, Manola, Horstmann, & Peters, 2010).
The linking that allows the creation of such models demands the identification of research entities, ideally using unambiguous identifiers. The most egregious of them:

- DOI, ARK - for publications (Freire, Manguinhas, Isaac, & Charles, 2023)
- ROR - for institutions (Welke & Krause, 2024)
- ORCID - for researchers (Schnieders et al., 2022)
- ISSN - for journals (Bequet, 2022)

And others that could be linked, such as patent identifiers, companies registries in the national offices, etc.
Usually databases of research information systems collect their data from forms manually filled by researchers and staff or by collecting other online databases. The diversity of sources and the intrinsic variable nature of human filled information creates challenges for matching entities across databases and inside the same databases. The variations in titles, in the writing of names given its multiple components, abbreviations and translated names for research institutions, all that adds to the importance of persistent identifiers.

Distance metrics to identify names

When matching of titles is needed, it is common to use metrics that compare pieces of text as strings (Slavin, Andreeva, & Putincev, 2022). These metrics use character variation as bases, the most common of them being the Levenshtein distance (Öztürk, Ertürk, Casale-Brunet, Ribeca, & Mattavelli, 2024; Sadiah, Iryani, Zuraiyah, Wahyuni, & Zaddana, 2024). Although widely used to compare human names (Kiawkaew, Kaothanthong, & Theeramunkong, 2023), it is not quite appropriate for this application, given that names are usually given in several alternative forms omitting family or given names or replacing them by initials.
An attempt to match names using such a metric would create artificial distances and proximities that make the matching less likely in many cases. To tend to these limitations, this work uses an adaptation of the Levenshtein metric on portions of the name, taking into account only names that appear and initials, also their relative order.

Data model for scientific production

Following the inspiration from OpenAlex and OpenAIRE (Manghi et al., 2010; Vichos et al., 2022) this paper adopts a relational model for the entities aiming at

their disambiguation. In this model, an author is linked to its publications by an authorship relation. The authorship contains an affiliation to an institution. Each work is connected to authorships as indicated in Figure 1.

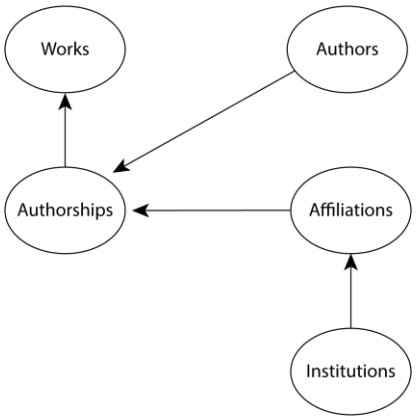


Figure 1. Data model for production.

Methodological Proposal and Rationale

Authors' identification

Given it is an official registry linked to the national identification code, it is safe to assume that the Lattes registry is unique and unambiguous about researchers. Starting from this, the matching proceeds to find them as authors in OpenAlex. The list publications provided for every author inside Lattes is not guaranteed to be complete, however, it is in the best interest of researchers to fill the list completely given that the evaluation from the national agencies are based on the Lattes information.

The publications metadata is manually filled and can contain errors and publications can be multiply registered by its coauthors. The strategy consists of selecting the publications with DOI for every author. Thus, several publications will share authors. That creates a list of possible authors from both sources (Lattes and OpenAlex) for every DOI.

Every list of possible authors is a matching pool for a comparison with a distance metric for names. That way, the number of authors to be compared is limited to a small number of coauthors. It is possible to assume that the most similar name is the correct matching if a certain distance threshold is assumed as the minimum acceptable. The improved metrics for names improves the possibility of finding author names. The threshold can be easily checked ordering the most likely author matching in each case from the worst distance to the best.

The method consists of finding a persistent identifier that collects an entity that must be matched in two databases. Using the smaller subset of possible matches filtered by the identifier, it is possible to evaluate against the distance metric for names.

In the case of the two databases here used, we find publications on the Lattes databases with explicit DOIs. There will be an unambiguous author associated with it. We want to locate this author inside OpenAlex. The found DOIs link to a list of coauthors of each publication. The coauthors are possible candidates of matching for the original author taken from Lattes. This way, the name matching happens on the smaller subset of possible coauthors. The best candidate according to the metric, given a certain threshold, is pointed as the match from Lattes to OpenAlex.
A point that should be observed is that, given a certain number of common DOIs between two candidate matching names, a certain threshold is appropriate to guarantee correct matching. However, with a larger number of DOIs in common, one could use a smaller threshold for the distance metric for names. That is so because more clues about the right candidate allow for a looser criterion for the matching. That way, the best threshold to use is a function of the number of common DOIs.
By checking the Lattes database, we can see that the largest possible number of common DOIs between authors' names is 649. We selected 0,1 distance (10% of length of name different from one to another) as the threshold for the worst possible case (just 1 DOI in common) and 0,4 (40% of difference) as the threshold for the situation with 649 DOIs in common. That leads to an inequality (with the threshold as t and the number of common DOIs as n):

$$t < \frac{64.5 + 0.3n}{648}$$

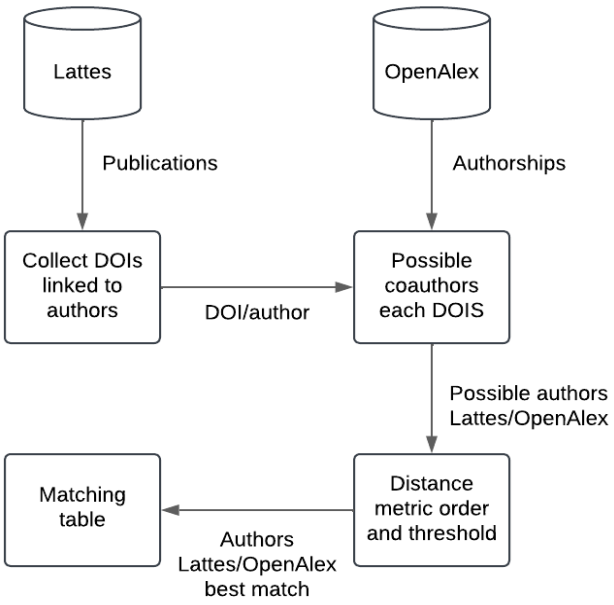


Figure 2. Data model for production.

The Lattes Database and its Application

In the late 1990s, Brazil's national funding agency recognized the need for a new approach to evaluating researchers' credentials. To address this, it first established a 'virtual community' comprising federal agencies and researchers to design and develop the Lattes infrastructure. This database provides high-quality data on approximately millions of researchers and thousands of institutions registered within it (Lane, 2010). Lane (2010) argues that the Brazilian experience with the Lattes Database (http://lattes.cnpq.br/english) exemplifies best practices in research assessment and creates appropriate incentives for researchers and academic institutions to utilize the database.

Results and Conclusion

We collected the data for all authors in Lattes that are PhDs and have published in the last 5 years. From their declared publications, we found all the ones with DOIs and applied our methodology.

The number of Lattes identifiers (representing researchers) with at least one valid DOI to apply our methodology is 154,474. The method identifies 151,318 authors in the OpenAlex database. Contrary to expected, this is not due to authors not being found, but to multiple people assigned to just one author in the OpenAlex database.

The number of matches with an exact correspondence of one to one is 148,431. This amounts to 6,043 people in Lattes being identified as 2,887 authors in OpenAlex. This strange phenomenon is mostly due to its disambiguation algorithm (Barrett, 2023) that compares names and takes fuzzy clues such as collaborations and areas of research in order to identify people. It relies ultimately on ORCID, but such an identifier is absent in many cases.

Figure 3 presents the number of DOIs used in the matching of the authors. Table 1 presents the number of incorrectly assigned to a single author identifier.

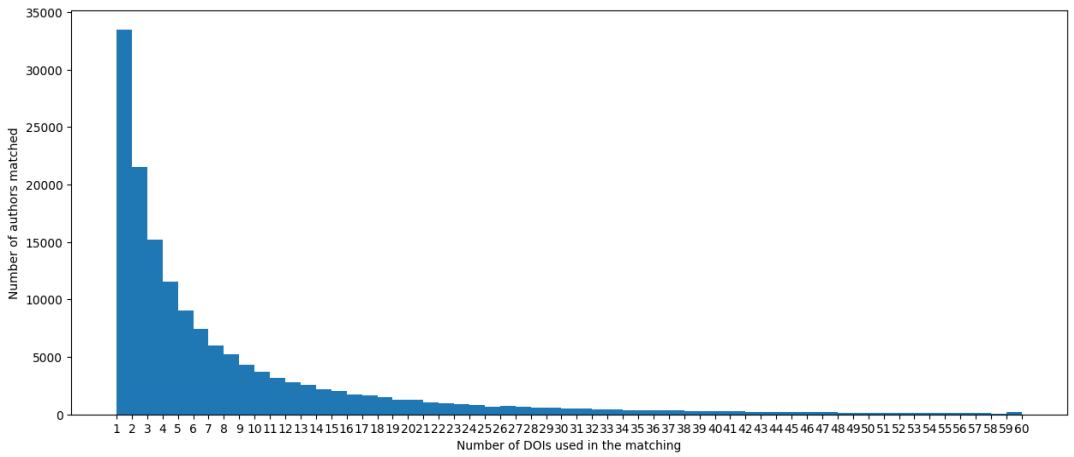


Figure 3. DOIs used in the matching of authors.

Table 1. Total of assigned identifiers to a single author identified after matches our methodology.

Total of identifiers	*Total of assigned*
1	148,431
2	2,645
3	220
4	17
5	5

The results show that the disambiguation algorithm has succeeded in a high percentage of the authors but plays on the risky side of assigning just one author identifier to a few people in some cases instead of a more conservative approach to keep doubtful cases split.

The use of the OpenAlex database however, given its open nature, allows for corrections such as the one here presented. A large deal of information can now be extracted for the authors that are uniquely identified. The smaller percentage of misidentified people can now be studied and the room for improvements based on local databases is paved. The methodology of combining persistent identifiers with an adapted metric in a dynamic threshold can be expanded to improve the database and its applications by adding other local databases, such as institutional data, for example.

References

Alshamaila, Y., Alsawalqah, H., Habib, M., Al-Madi, N., Faris, H., Alshraideh, M., Aljarah, I., et al. (2024). An intelligent rule-oriented framework for extracting key factors for grants scholarships in higher education. *International Journal of Data and Network Science*, *8*(2), 1325–1340.

Barrett, J. (2023). Openalex name disambiguation. Retrieved from https://github.com/ourresearch/openalex-name-disambiguation/tree/main/V3

Bequet, G. (2022). From the Cradle to the Digital Vault: Tracking the Path of E-journals. *Serials Librarian*, *82*(1–4), 199–204.

Boyle, F., & Sherman, D. (2006). Scopus™: The Product and Its Development. *The Serials Librarian*, *49*(3), 147–153.

de Castro, P., & Puuska, H.-M. (2023). Research Information Management Systems: Covering the whole research lifecycle (Vol. 95, pp. 257–265). Presented at the EPiC Series in Computing.

Digiampietri, L. A., Mena-Chalco, J. P., Pérez-Alcázar, J. J., Tuesta, E. F., Delgado, K. V., Mugnaini, R., & Silva, G. S. (2012). Minerando e Caracterizando Dados de Currículos Lattes. Retrieved from https://sol.sbc.org.br/index.php/brasnam/article/view/6868

Ferreira, A. A., Gonçalves, M. A., & Laender, A. H. F. (2012). A brief survey of automatic methods for author name disambiguation. *ACM SIGMOD Record*, *41*(2), 15–26.

Freire, N., Manguinhas, H., Isaac, A., & Charles, V. (2023). Persistent Identifier Usage by Cultural Heritage Institutions: A Study on the Europeana.eu Dataset. In O. Alonso, H. Cousijn, G. Silvello, M. Marrero, C. Teixeira Lopes, & S. Marchesin (Eds.), *Linking Theory and Practice of Digital Libraries* (pp. 341–348). Cham: Springer Nature Switzerland.

Kiawkaew, T.-A., Kaothanthong, N., & Theeramunkong, T. (2023). A Practical Technique for Thai-English Word Mapping Using Phonetic Rules: Person Name Matching Case Study. *2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1–6). Presented at the 2023 18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). Retrieved January 25, 2025, from https://ieeexplore.ieee.org/document/10354663/?arnumber=10354663

Kim, J., & Kim, J. (2018). The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics*, *117*(1), 511–526.

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology*, *63*(5), 1030–1047.

Manghi, P., Manola, N., Horstmann, W., & Peters, D. (2010). An Infrastructure for Managing EC Funded Research Output – The OpenAIRE Project –.

Mena-Chalco, J. P., & Junior, R. M. C. (2009). scriptLattes: An open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society*, *15*(4), 31–39. SpringerOpen.

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, *106*(1), 213–228.

Öztürk, Ü., Ertürk, U. G., Casale-Brunet, S., Ribeca, P., & Mattavelli, M. (2024). Efficient Neural Clustering and Compression of Strings Through Approximate Euclidean Embeddings of the Levenshtein Distance. *2024 Data Compression Conference (DCC)* (pp. 575–575). Presented at the 2024 Data Compression Conference (DCC). Retrieved January 25, 2025, from https://ieeexplore.ieee.org/document/10533750/?arnumber=10533750

Qian, Y., Hu, Y., Cui, J., Zheng, Q., & Nie, Z. (2011). Combining machine learning and human judgment in author disambiguation. *Proceedings of the 20th ACM international conference on Information and knowledge management* (pp. 1241–1246). Presented at the CIKM '11: International Conference on Information and Knowledge Management, Glasgow Scotland, UK: ACM. Retrieved January 20, 2025, from https://dl.acm.org/doi/10.1145/2063576.2063756

Rehs, A. (2021). A supervised machine learning approach to author disambiguation in the Web of Science. *Journal of Informetrics*, *15*(3), 101166.

Sadiah, H. T., Iryani, L. D., Zuraiyah, T. A., Wahyuni, Y., & Zaddana, C. (2024). Implementation of Levenshtein Distance Algorithm for Product Search Query Suggestions on Koro Pedang Edutourism E-Commerce. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, *42*(2), 188–196.

Sanyal, D. K., Bhowmick, P. K., & Das, P. P. (2021). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, *47*(2), 227–254.

Schnieders, K., Mierz, S., Boccalini, S., Meyer zu Westerhausen, W., Hauschke, C., Hagemann-Wilholt, S., & Schulze, S. (2022). ORCID coverage in research institutions—Readiness for partially automated research reporting. *Frontiers in Research Metrics and Analytics*, *7*.

Simón, L., Fontáns, E., & Aguirre-Ligüera, N. (2013). El currículum vitae como fuente de datos en los estudios métricos. Retrieved from http://sedici.unlp.edu.ar/handle/10915/38075

Slavin, O., Andreeva, E., & Putincev, D. (2022). Application of modified Levenshtein distance for classification of noisy business document images. *Fourteenth International Conference on Machine Vision (ICMV 2021)* (Vol. 12084, pp. 78–85). Presented at the Fourteenth International Conference on Machine Vision (ICMV 2021), SPIE. Retrieved January 25, 2025, from https://www.spiedigitallibrary.org/conference-proceedings-of-

spie/12084/120840B/Application-of-modified-Levenshtein-distance-for-classification-of-noisy-business/10.1117/12.2623437.full

Turgel, I. D., & Chernova, O. A. (2024). Open Science Alternatives to Scopus and the Web of Science: A Case Study in Regional Resilience. *Publications*, *12*(4), 43. Multidisciplinary Digital Publishing Institute.

Vichos, K., De Bonis, M., Kanellos, I., Chatzopoulos, S., Atzori, C., Manola, N., Manghi, P., et al. (2022). A Preliminary Assessment of the Article Deduplication Algorithm Used for the OpenAIRE Research Graph (Vol. 3160). Presented at the CEUR Workshop Proceedings.

Welke, B., & Krause, B. (2024). Automatically generated Research Profiles for Experts, Institutions and Working Groups (Vol. 249, pp. 112–119). Presented at the Procedia Computer Science.

Xu, J., Shen, S., Li, D., & Fu, Y. (2018). A Network-embedding Based Method for Author Disambiguation. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (pp. 1735–1738). Presented at the CIKM '18: The 27th ACM International Conference on Information and Knowledge Management, Torino Italy: ACM. Retrieved January 20, 2025, from https://dl.acm.org/doi/10.1145/3269206.3269272

https://doi.org/10.51408/issi2025_162

Ages in Academia: How Faculty Age Shapes University Research Output

Anastasia Byvaltseva-Stankevich[1], Anna Panova[2]

[1]*abyvaltseva@hse.ru*, [2]*apanova@hse.ru*
National Research University Higher School of Economics (HSE University), Pokrovsky Blvd. 11, 109028 Moscow (Russian Federation)

Abstract

Understanding the influence of faculty age structure on research productivity is crucial for university strategy, particularly amidst workforce ageing. This study investigates how the age structure of faculty affects research productivity of Russian universities, focusing on both quantitative and qualitative outcomes. We use the data on over 300 state universities from 2014 to 2020 and estimate a panel negative binomial regression with instrumental variables and fixed time effects. The results indicate that the relationship between faculty age and productivity differs based on the university type – whether it belongs to the group of leading (research-intensive) universities or not, which indicates the importance of the environment where faculty work. At non-leading universities, young faculty make the greatest contribution to university research productivity, particularly when mostly national journals are considered, while at leading universities older faculty outperform their younger colleagues, especially in terms of internationally recognized publications and their citations.

Introduction

Apparently, age demographics can affect organizational effectiveness, yet the details of this relationship may vary across countries, industries and institutional environments (Grund & Westergaard-Nielsen, 2008; Frosch, 2011). The age structure of faculty within universities can influence their organizational performance, much like in other sectors. Nowadays universities are a significant part of the global research community, and conducting research is one of the duties expected from almost any modern university faculty member. Modern universities, especially leading ones, strive to enhance their research output which influences their local and international academic reputation. Quite often research performance is prioritized as a metric of university success in global rankings and national universities excellence initiatives (Salmi, 2016). Thus, understanding how the demographics of faculty members influences their general publishing activity becomes crucial for universities. Based on these insights and mission statement, universities can adjust their policies regarding faculty recruitment and motivation.

A global trend that is evident in education is the workforce ageing. The increase in the average age of university faculty is relevant for many countries (Earl, Taylor, & Cannizzo, 2017; Finkelstein, 2008; Stein, 2000), and Russia is no exception (Byvaltseva-Stankevich & Panova, 2025). Although the question of faculty ageing is discussed in the literature, the ways it affects faculty productivity remains unaddressed for universities. In our paper we investigate how the age structure of faculty contributes to Russian universities' research productivity.

Background

The research productivity of people of different ages can be analysed from two perspectives. First, the direct impact of age can be discussed. Age influences an individual's cognitive abilities, knowledge base, and professional networks, which can evolve throughout their career and consequently affect research output. The second perspective focuses on the influence of the cohort to which a researcher belongs. The cohort unites people who belong to the same generation, and therefore share similar educational experiences, career opportunities, and prevailing scientific paradigms that define their professional trajectory.

The research productivity of different ages is discussed and analysed in the literature. While there may be a tendency for behavioural slowing as individuals become older, the experience and wisdom of senior academics make them particularly important for the research community (Birren, 1990). Another theoretical concept suggests that researchers may intentionally decrease their research activity as they age, since they may find greater value in pursuing alternative activities (Kwiek, 2015). In line with ageing, faculty become involved in a greater range of tasks not directly related to their own research, and the limited nature of time resources available suggests that older faculty will devote less time to research and therefore become less productive in this sphere (Costas, van Leeuwen, & Bordons, 2010; Kwiek, 2015).

Empirical results regarding the link between age and productivity of individual researchers are also controversial. According to (Kyvik, 1990), there is a curvilinear relationship between age and productivity of Norwegian faculty, with publishing activity reaching a peak during the ages of 45-49. The data on Polish faculty show almost the same results (Kwiek, 2018): the mean age of top performers in terms of research productivity is around 50, varying across different fields of study. Polish faculty decrease time spent on research for the sake of teaching (Kwiek, 2015). Malaysian data suggests that the "golden years" of an academic's productivity are around forties, but might be different, depending on the model used to measure this (Yen, Lim, & Campbell, 2015). In one case, there is an inverted-U shaped relationship with the peak at 41 years; another model shows that faculty have two most productive periods in their careers – when they are 34-40 years old and between ages 46-50. Results on the American data suggest that in general older scientists are less productive that younger ones. While there might be an increase in performance at the ages of 40-55, depending on the research sphere, younger faculty members outperform their older peers in terms of papers published (Levin & Stephan, 1989, 1991). Nevertheless, the latter authors claim that age might be a weak predictor of research performance. Negative relationship between age and research productivity was found for the UK psychologists working at universities (Over, 1982) and for Italian full professors (Abramo, D'Angelo, & Murgia, 2016). Therefore, the question of research productivity of different ages remains open.

Another perspective on the relationship between academic productivity and faculty demographics is the cohort effect, which suggests that generational differences in research performance may be explained not only by individual age, but also by the shared experiences of faculty members who entered academia at a particular time

(Stephan & Levin, 1992). Unlike age-related declines in productivity, which are related to cognitive and behavioural changes, cohort effects are driven by external conditions that shape academic careers. These can include changes in funding policies, institutional incentives, technological advances and global trends in higher education. Research productivity of different cohorts was previously studied on data from several countries. Results for Norwegian academic staff suggest that potential reasons why current young generations perform at a higher level might be better qualification, new incentive systems and norms of academic behaviour, improved funding and research conditions, changing patterns in research collaboration and young generation's readiness to get involved in it (Kyvik & Aksnes, 2015; Kyvik & Olsen, 2008). Similar conclusions were made by comparing two cohorts of Spanish researchers: differences in publication culture and incentive systems result in differences in productivity across cohorts (Albert, Davia, & Legazpe, 2016).

The university brings together people of different ages, creating environment where diverse experiences enrich research. It is not known how universities with different age structures succeed, since previous studies have been conducted on the individual level. The age structure of faculty can affect university productivity, as different age groups may contribute to research in different ways. An optimal age balance ensures sustainable university development, mitigates staffing risks and increases research effectiveness.

Russian higher education sector has also faced the challenge of the ageing workforce: the average share of faculty under the age of 40 at Russian universities has decreased from 2014 and 2020.

The cohort effect is particularly relevant in the Russian context, where the academic environment has changed significantly in recent decades. Faculty members who started their careers in the 1990s faced severe underfunding, brain drain and weak research infrastructure, which could have long-term effects on their productivity. In contrast, researchers who entered academia in the 2010s were exposed to a highly competitive environment driven by publication-oriented policies and international rankings. These systemic differences suggest that research productivity may not only vary with age but also depend on the specific career trajectories of different academic cohorts.

Nevertheless, the question of research productivity of different age cohorts has not been tested on the Russian data. In line with institutional changes described below, this might have influenced the research productivity of universities.

The Russian university system has inherited several characteristics from the Soviet times, including a generally low engagement in research activities and a contract-based employment system. Historically, only a limited number of universities, such as Moscow State University and Saint Petersburg State University, were actively involved in research. In 2000-s the government made several comprehensive steps towards integrating research in the sector of higher education. Some universities were awarded the status of federal universities, following by the designation of the national research university status for select universities. The latter status implied receiving additional funding for staff development, the purchase of new equipment,

and the improvement of research systems. Another major attempt to stimulate research at universities was made in 2012, when the Russian university excellence initiative (RUEI) was launched. This project selected the best universities in Russia with the intention to increase their research and teaching to a level that can compete with the world leaders. Universities and their progress were assessed by a set of criteria with the research productivity being the most important, which implied that a university's success in the program was determined by the number of papers published by its faculty, especially in international and high-quality journals, and their citations. Even though all universities participating in the program had the same standards to aspire to and received substantial financial support, they were not equally successful in terms of publishing activity. Some participants made substantial progress and climbed up in international rankings (Volkov, Kuzminov, & Yudkevich, 2023). Owing to productive universities participating in the program, the university sector has managed to reach the research sector in terms of publications or even outperform it in some disciplines (Lovakov & Panova, 2023). Russian universities that have not taken part in the program also differ substantially in their research productivity. While some of them are close to the level of leading universities, faculty of other universities scarcely publish research papers in peer-reviewed journals.

Additional peculiarities of the Russian system include low mobility among researchers and the prevalence of fixed-term contracts, contrasting with tenure-track positions common in other countries (Panova & Yudkevich, 2021). Recent reforms include the implementation of performance monitoring for universities, aimed at addressing inefficiencies, and the introduction of targeted funding programs to boost research activities in select universities. These initiatives have contributed to the formation of a distinct sector of leading research universities and the widespread adoption of performance-based contracts.

Now there is a strong differentiation of Russian universities. Two clusters of universities – leading and non-leading universities – differ in their goals, policies and orientation. The cluster of leading universities, which mostly consists of the RUEI participants, is research-oriented and concerned about international standing, while other universities are not as concerned about publications and rankings. Such differentiation of universities' goals is linked to different working environments and reward systems created there, which might also have an impact on the relationship between age and research performance (Cole, 1979).

Thus, Russian higher education sector and its historical background have peculiarities. Our research fills a gap by providing insights into the Russian higher education landscape accounting for its context. This paper investigates the relationship between the distribution of faculty into different age cohorts and publishing activity of Russian universities. In contrast to the previous studies, we use the university-level data and focus on the organizational productivity rather than on individual faculty members. Thus, the aim of our research is to estimate how the age structure of faculty influences Russian universities' research productivity in terms of quantity – the number of papers published, and in terms of quality – the citations of

the papers published. Moreover, we measure productivity with different types of publications, depending on the citation databases in which they appeared, which indicates, among other things, the quality of papers and orientation of universities.

Analysis

Data and variables

We analyse the research performance of universities in terms of quantitative productivity and qualitative productivity. The analysis is based on the data reported annually by universities and collected by the Monitoring of University Efficiency. To measure university's research productivity from the quantitative perspective, we use the number of papers (per 100 faculty) published by its faculty within a year in the journals indexed in three different citation databases: Web of Science (WoS), Scopus, Russian Science Citation Index (RSCI). To measure university's research productivity from the qualitative perspective, we use the number of citations (per 100 faculty) of papers published by its faculty over the last five years in the journals from the same three databases. WoS and Scopus journals represent mostly international and high-quality journals, while RSCI includes mostly Russian journals of a relatively lower quality in comparison to WoS and Scopus (Kassian & Melikhova, 2019), therefore analysis of papers and citations of papers from different databases will give more insights into the quality of publications and universities' orientation – national or international. The data on the number of papers published by university faculty and indexed in different databases in provided in the aggregated format by the Monitoring of University Efficiency.

Explanatory variables of interest are the percent of faculty under the age of 40, namely young faculty, and the share of faculty older than 65 years of age, namely old faculty. The remaining part are the middle-aged faculty members, which are not included explicitly due to perfect multicollinearity, thus this group will be the reference category while discussing the results. We also include the interaction terms between the variables of interest and the university status – whether it belongs to the group of leading universities or not. Leading universities include the Russian university excellence initiative (RUEI) participants and two more universities with a special status and similar goals. Leading universities are research intensive, therefore research patterns may differ there, in comparison to non-leading universities. What is more, this helps us to account for the fact that the RUEI, among other things, included such measures as attracting and supporting young faculty, improving postgraduate and doctoral studies. We also control for various university characteristics. To account for the fact that publication patterns may differ depending on the field considered, we include the set of profile variables: even though we determine university's profile according to its teaching major, this is a relevant proxy for university's research orientation since Russian universities tend to publish in the fiends of science related to their teaching orientation (Tsivinskaya, 2023). The full set of variables used in the analysis is presented in Table 1.

Table 1. Variables description.

Variable name	*Variable meaning*
Productivity (dependent variables):	
WoS / Scopus / RSCI papers	number of papers per 100 faculty published by university's faculty within a year in the journals indexed in WoS, Scopus or RSCI, respectively
WoS / Scopus / RSCI citations	number of citations per 100 faculty of papers published by university's faculty over the last five years in the journals indexed in WoS, Scopus or RSCI, respectively
Variables of interest:	
percent of young / old	percent of young faculty (under the age of 40) and old faculty (older than 65 years of age), respectively
Control university characteristics:	
leading status	= 1 if a university is from the group of leading universities
offline students	number of students (in thousands) studying offline as a proxy for the university size
students per faculty	number of students per faculty member as a measure of faculty average teaching load
share of foreign	share of foreign faculty
PhD students	number of PhD students per 100 bachelor and master students
hard profile	= 1 if university's major teaching profile is related to hard sciences (technologies, physics, engineering, informatics, geodesy, electronics, math, nucleonics, mechanics, metallurgy, etc.)
agriculture profile	= 1 if university's major teaching profile is related to agriculture, forestry or fisheries
medicine profile	= 1 if university's major teaching profile is related to medicine or pharmacology
art & sport profile	= 1 if university's major teaching profile is related to art or sports

Econometric model

We estimate the effect of the age structure on universities' research productivity using negative binomial panel data regressions for over 300 Russian state

universities in the period 2014-2020 with fixed time effects. We do not include fixed individual effects in order to be able to estimate the impact of those variables that do not change during the period of study. Negative binomial regression is used to account for the left-hand skewness of the dependent variables since there are many universities with very low numbers of papers and citations.

We also account for endogeneity occurring because of the potential simultaneity. Since university's publishing activity may give signals to candidates about the potential workload related to research, the shares of young and old faculty may depend on those indicators that are used as dependent variables. This is especially true for the young faculty whose desire to work at a certain university might be influenced by the potential necessity to do research and publish papers along with teaching (Byvaltseva-Stankevich & Panova, 2025).

Thus, as the first stage of the econometric modelling we estimate regressions of the percentages of young and old faculty in the current year on their time lags as instruments and on exogenous variables. After instrumenting the shares of young and old faculty with their values in the previous period, we insert the first-stage estimation results in the negative binomial regression to cope with endogeneity and get consistent coefficient estimates on the second stage. Here we also include the interaction terms between the shares of young and old faculty and the university status to account for potential differences in the effects we are estimating. The formal description of the estimation procedure is outlined below in Box 1.

Box 1. Estimation procedure.

Stage 1:

$$\begin{cases} Young_{it} = \alpha + \gamma\, Young_{i,t-1} + X_{it}\beta + \varepsilon_{it} \\ Old_{it} = \mu + \delta\, Old_{i,t-1} + X_{it}\theta + \epsilon_{it} \end{cases},$$

where *i* is the university indicator,
t is the year indicator,
Young and *Old* are the percentages of young and old faculty, respectively,
X is the set of exogenous variables (apart from the variables themselves, includes the interaction term, between the leading status and the students per faculty indicator),
$\alpha, \gamma, \mu, \delta$ are coefficients and β, θ are vectors of coefficients being estimated,
ε, ϵ are error terms.
These regressions are estimated using the ordinary least squares method. After estimation, predictions of the percentages of young and old faculty are obtained:

$$\begin{cases} \widehat{Young}_{it} = \hat{\alpha} + \hat{\gamma}\, Young_{i,t-1} + X_{it}\hat{\beta} \\ \widehat{Old}_{it} = \hat{\mu} + \hat{\delta}\, Old_{i,t-1} + X_{it}\hat{\theta} \end{cases}$$

Stage 2:

$$\mathbb{P}(Productivity_{it}) = F(\eta + \lambda\, \widehat{Young}_{it} + \varphi\, \widehat{Old}_{it} + X_{it}\xi + \\ + \vartheta\, \widehat{Young}_{it} * leading_i + \psi\, \widehat{Old}_{it} * leading_i + \tau_t),$$

where $\widehat{Young}$ and $\widehat{Old}$ are the first-stage predictions of percentages of young and old faculty, respectively,
$\eta, \lambda, \varphi, \vartheta, \psi$ are coefficients and ξ is the vectors of coefficients being estimated, τ are the fixed time effects being estimated (effects of years 2016-2020 are estimated explicitly, 2015 is the base reference year, year 2014 is not estimated since it was used for lags at Stage 1),
F(.) is the probability function for the negative binomial model, which is the Poisson-gamma mixture,
other notations remain unchanged.
Regressions (separate regressions for different productivity indicators) are estimated using the maximum likelihood method. Coefficients estimates are then exponentiated for calculating incidence rates that can be interpreted in a more intuitive way. Thus, the effects of young and old faculty groups at non-leading universities are $e^{\hat{\lambda}}$ and $e^{\hat{\varphi}}$, respectively. At leading universities, the effects of young and old faculty groups are cumulative: $e^{\hat{\lambda}+\hat{\vartheta}}$ for the youngest group and $e^{\hat{\varphi}+\hat{\psi}}$ for the oldest group.

Results

The estimated regression allows for the identification of the contribution of each age group, relative to the middle-aged group that was taken as the reference group, to the overall productivity of a university. Since negative binomial regression coefficient

estimates show the change in the logarithm of the dependent variable, it is more intuitive to interpret the incidence rates estimates. They are calculated as the exponential function of the coefficient estimates and therefore, show the change in dependent variable in case the corresponding explanatory variable rises by one unit. In case the incidence rate is higher than 1, this means the rise of the dependent variable, while the value below 1 indicates the drop of the dependent variable. The full sets of coefficients and incidence rates estimates are provided in the appendices (Appendix 1 and Appendix 2).

Here we focus on the variables of interest, and Table 2 below shows the effects of young and old cohorts at non-leading and leading universities. Statistical significance of the incidence rates at non-leading universities is determined by statistical significance of the corresponding coefficients. For the group of leading universities, the effect is cumulative (the details of calculation are provided in the Box 1), and its statistical significance is tested manually by testing whether the sum of corresponding coefficients is equal to zero.

Table 2. The effects of age groups (exponentiated coefficients for non-leading universities, exponentiated sums of coefficients for leading universities).
Source: authors' calculations.

		Quantitative productivity (number of papers)			*Qualitative productivity (citations of papers)*		
university status	*age group*	*WoS*	*Scopus*	*RSCI*	*WoS*	*Scopus*	*RSCI*
non-leading	young	1.001	1.012^{***}	1.013^{***}	0.972^{*}	0.984	1.012^{***}
	old	1.000	1.007^{**}	0.986^{***}	0.966^{**}	0.979^{*}	0.988^{***}
leading	young	1.027^{*}	1.031^{**}	1.002	0.955^{**}	0.977	1.021^{*}
	old	1.048^{***}	1.045^{***}	0.996	1.034^{*}	1.051^{***}	1.053^{***}

Note: Stars correspond to p-values: $^{*}p<0.1$, $^{**}p<0.05$, $^{***}p<0.01$.
Bootstrapped standard errors were used for p-values calculation.

To have more intuitive interpretation for further discussion of the results, we can say that the rise of the explanatory variable by one unit is associated with the change of the dependent variable by (incidence rate – 1) * 100%. Here we provide another table that shows only significant changes, and these changes are presented in percentages. Thus, numbers in Table 3 provided below show what happens with the dependent variable in case the explanatory variable rises by 1.

Table 3. Significant changes in the dependent variable.
Source: authors' calculations.

		Quantitative productivity (number of papers)			*Qualitative productivity (citations of papers)*		
university status	*age group*	*WoS*	*Scopus*	*RSCI*	*WoS*	*Scopus*	*RSCI*
non-leading	young		↑ 1.2%	↑ 1.3%	↓ 2.8%		↑ 1.2%
	old		↑ 0.7%	↓ 1.4%	↓ 3.4%	↓ 2.1%	↓ 1.2%
leading	young	↑ 2.7%	↑ 3.1%		↓ 4.5%		↑ 2.1%
	old	↑ 4.8%	↑ 3.8%		↑ 3.4%	↑ 6.5%	↑ 5.3%

Based on these results, we define the most and the least productive age cohorts with the productivity measured in terms of quantity or quality. In case the effect of a certain group was not statistically significant, this means that the productivity of this age group does not differ from the productivity of the reference middle-aged group. Table 4 below provides information on the most and least productive groups at both types of universities.

Table 4. Most and least productive age groups.

		Quantitative productivity (number of papers)			*Qualitative productivity (citations of papers)*		
university	*productive group*	*WoS*	*Scopus*	*RSCI*	*WoS*	*Scopus*	*RSCI*
non-leading	most		young	young	middle		young
	least		middle	old	old	old	old
leading	most	old	old		old	old	old
	least	middle	middle		young		middle

Note: Empty cell means that the productivity of the remaining groups has no statistically significant difference

The results on Russian data differ from those that were previously obtained for other countries when the productivity was measured on the individual level. For such countries as Norway (Kyvik, 1990), Poland (Kwiek, 2018), Malaysia (Yen et al., 2015) and the USA (Levin & Stephan, 1989, 1991), there was found an inverted U-shaped relationship between age and productivity. However, our results do not correspond with this type of relationship since middle-aged group is never the most productive group, with only one exception. Moreover, our results suggest that this

relationship is different for universities' research productivity measured by papers or citations of papers from different categories.

If we consider non-leading universities, the youngest faculty cohort appears to be the most productive in most cases there. As for WoS papers and Scopus citations, this age cohort's productivity does not differ significantly from the productivity of their middle-aged colleagues. However, the youngest cohort is outperformed by their middle-aged colleagues at non-leading universities when WoS citations are considered. The oldest cohort is often the least productive at non-leading universities.

On the contrary, the situation is different at leading universities where the oldest cohort is associated with the highest productivity, especially when it is measured qualitatively. The oldest cohort especially outperforms their middle-aged colleagues who are sometimes the least productive at selective universities. Similarly to the situation at non-leading universities, the youngest cohort is not productive relative to their colleagues of other ages when WoS citations are considered. When papers in RSCI journals are considered, all three age groups are similar in terms of their productivity.

The overall productivity scheme described above differs a lot at different types of universities: while at leading universities the oldest cohort is generally the most productive, at non-leading universities is the youngest group that outperforms its peers of other ages. Such disparities, in combination with the unchanged relative productivity between young and middle-aged groups, indicate that the contribution of age groups to the overall productivity depends on the environment they are working in.

Discussion and conclusion

Our study examines how the age structure of faculty influences university research productivity by comparing the relative contributions of three age cohorts. The results indicate that the relationship between age structure and productivity is not universal but depends on the type of university. While younger faculty tend to contribute more at non-leading universities, the highest contribution at leading universities is generated by the oldest cohort. These patterns suggest that faculty productivity is shaped not only by individual career stages but also by the incentives, traditions, and policies of the universities they work in.

Russian higher education system has been through various stages, which is still echoed in its current practices and faculty. The oldest cohort in our data, who are aged 65 and older, got their education in the Soviet times. Moreover, most likely, their academic careers started in the Soviet times and faced the crisis associated with the USSR breakdown. Their careers started in a system when most universities were not actively involved in research activities. The middle-aged cohort includes those individual whose education or early career was associated with this crisis, while the youngest cohort, who are less than 40, got their higher education after the crisis period, when the government established new higher education system. The post-soviet crisis led to the decline in the higher education financing, that is why it was

likely that the most talented faculty of those period decided to leave or to switch to more prestige universities. On the contrary, faculty with the lower productivity remained at their universities after the crisis. Redistribution of the talent during the crisis period and reforms might have left their imprint on the current productivity of the oldest cohort. While they are not productive at non-leading universities, this age cohort is the most productive at leading universities. The combination of factors – academic tradition expressed in involvement in research, talent concentration, and the stimuli provided by leading universities, such as monetary rewards for publications in top journals, – enables the oldest cohort to conduct fruitful research at leading universities. The only result that seems contradictory is the fact that the oldest cohort does not outperform their younger colleagues at leading universities when papers in RSCI journals are considered. However, this might be explained by the fact that older workers have enough skills and experience and are ready to invest their time and effort in long-term projects, which are WoS and Scopus papers, that is why papers in top journals are prioritized over RSCI papers. Older faculty at non-leading universities are unmotivated and, as a result, are not publishing or being cited, nor are they investing in long-term, high-quality projects. This suggests that well-established research environments and appropriate institutional stimuli can sustain high productivity even at later career stages.

Discussing the relative productivity of the youngest cohort, in most cases they are the most productive group at non-leading universities. It seems that at non-leading universities the youngest cohort is the most motivated since these faculty members include those who are in the process of writing their PhD theses, which requires publications. Other faculty members do not have strong incentive to publish since non-leading universities do not stimulate them and the most talented older faculty are employed by leading universities. The only indicator, according to which the youngest cohort is outperformed by their middle-aged colleagues at non-leading universities is WoS citations. WoS journals are highly selective and gaining enough experience to publish there and get citations takes time, but young faculty have not had enough time for accumulating enough experience and citations thus far, therefore their productivity measured by WoS citations is not the highest among faculty.

Another interesting result is the fact that compelling distinctions between the productivity schemes at two types of universities are combined with the unchanged relative productivity between young and middle-aged groups. This allows us to say that the contribution of cohorts to productivity depends on the environment they are working in. Academic environment includes colleagues, incentives created by universities, academic tradition prevailing at a university. Since these factors differ at leading and non-leading universities, this is reflected in the disparities in the link between age and productivity at different types of universities. This once again emphasises that leading universities have developed a favourable academic environment, unlike non-leading universities.

Understanding the link between faculty age structure and university research productivity is essential for institutions aiming to enhance their research output and academic reputation. The results of this study can guide university administrators in

formulating targeted hiring strategies that balance the age distribution of faculty and in developing motivational practices and support systems focused on certain age groups. Depending on the goals of the state and universities, they could either help the most productive groups to maintain the same level of productivity or stimulate the least productive age cohorts to correspond to their colleagues of other ages. The latter decision will need further research because it is important to determine the reasons for such disparities between different age cohorts. This may have to deal with their physical and psychological abilities, as well as with institutional conditions and opportunities.

The major limitation of the study is the data available for research. The Monitoring of University Efficiency, which is the only public source providing information on Russian universities, categorizes faculty into specific age groups. Thus, we are limited to these groups and do not have an opportunity to alter the thresholds. Nevertheless, such age breaks are quite natural from the two points of view. First, they correspond to the way age is received in Russia: according to the state legislation, young researchers are those who are aged under 39 inclusive, while 65 years is the lower boarder of the retirement age. Secondly, age breaks used in the analysis allow for the clear differentiation of three cohorts that had different experience in terms of getting their education and developing their academic careers. Despite this, classification of faculty by broad age groups restricts a more granular analysis. Future research should explore individual-level trajectories and institutional strategies that enhance faculty engagement in research throughout their careers.

By highlighting the interplay between academic age structures and institutional environments, this study contributes to the broader discussion on how universities can develop sustainable research strategies in a rapidly changing higher education landscape.

References

Abramo, G., D'Angelo, C. A., & Murgia, G. (2016). The combined effects of age and seniority on research performance of full professors. *Science and Public Policy, 43*(3), 301–319. Oxford University Press.

Albert, C., Davia, M. A., & Legazpe, N. (2016). Determinants of Research Productivity in Spanish Academia. *European Journal of Education, 51*(4), 535–549.

Birren, J. E. (1990). Creativity, productivity, and potentials of the senior scholar. *Gerontology & Geriatrics Education, 11*(1–2), 27–44.

Byvaltseva-Stankevich, A., & Panova, A. (2025). Application of spatial econometric methods to analyze factors attracting young faculty to Russian universities. *Applied Econometrics, 77,* 91–115.

Cole, S. (1979). Age and scientific performance. *American journal of sociology, 84*(4), 958–977. University of Chicago Press.

Costas, R., van Leeuwen, T. N., & Bordons, M. (2010). A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact. *Journal of the American Society for Information Science and Technology, 61*(8), 1564–1581.

Earl, C., Taylor, P., & Cannizzo, F. (2017). "Regardless of age": Australian university managers' attitudes and practices towards older academics. *Work, Aging and Retirement, 4*(3), 300–313.
Finkelstein, M. (2008). Recruiting and retaining the next generation of college faculty: Negotiating the new playing field. In D. A. Heller & M. B. d'Ambrosio (Eds.), *Generational shockwaves and the implications for higher education* (pp. 82–100). Cheltenham: Edward Elgar Publishing.
Frosch, K. H. (2011). Workforce age and innovation: A literature survey. *International journal of management reviews, 13*(4), 414–430.
Grund, C., & Westergaard-Nielsen, N. (2008). Age structure of the workforce and firm performance. *International Journal of Manpower, 29*(5), 410–422.
Kassian, A., & Melikhova, L. (2019). Russian Science Citation Index on the WoS platform: A critical assessment. *Journal of Documentation, 75*(5), 1162–1168. Emerald Publishing Limited.
Kwiek, M. (2015). Academic generations and academic work: Patterns of attitudes, behaviors, and research productivity of Polish academics after 1989. *Studies in Higher Education, 40*(8), 1354–1376. Taylor & Francis.
Kwiek, M. (2018). High research productivity in vertically undifferentiated higher education systems: Who are the top performers? *Scientometrics, 115*(1), 415–462.
Kyvik, S. (1990). Age and scientific productivity. Differences between fields of learning. *Higher Education, 19*(1), 37–55.
Kyvik, S., & Aksnes, D. W. (2015). Explaining the increase in publication productivity among academic staff: A generational perspective. *Studies in Higher Education, 40*(8), 1438–1453. SRHE Website.
Kyvik, S., & Olsen, T. (2008). Does the aging of tenured academic staff affect the research performance of universities? *Scientometrics, 76*(3), 439–455.
Levin, S. G., & Stephan, P. E. (1989). Age and research productivity of academic scientists. *Research in Higher Education, 30*, 531–549.
Levin, S. G., & Stephan, P. E. (1991). Research productivity over the life cycle: Evidence for academic scientists. *The American economic review*, 114–132. JSTOR.
Lovakov, A., & Panova, A. (2023). The contribution of universities to the production of basic scientific knowledge in Russia. *Herald of the Russian Academy of Sciences, 93*(4), 221–230. Springer.
Over, R. (1982). Does research productivity decline with age? *Higher Education, 11*(5), 511–520.
Panova, A., & Yudkevich, M. (2021). Research and Higher Education in Russia: Moving closer together. *Universities in the knowledge society: The nexus of national systems of innovation and higher education* (pp. 183–201). Springer.
Salmi, J. (2016). Excellence initiatives to create world-class universities: Do they work. *Higher Education Evaluation and Development, 10*(1), 1–29.
Stein, D. (2000). Age and the university workplace: A case study of remaining, retiring, or returning older workers. *Human Resource Development Quarterly, 11*(1), 61–80.
Stephan, P. E., & Levin, S. G. (1992). *Striking the mother lode in science: The importance of age, place, and time*. Oxford University Press.
Tsivinskaya, A. (2023). The diversity of university disciplinary profiles in research and teaching. *Higher Education Quarterly, 77*(4), 853–873. Wiley Online Library.
Volkov, A., Kuzminov, Y., & Yudkevich, M. (2023). A project for the elite that changed the system as a whole: A case study of project 5–100. *Academic Star Wars: Excellence Initiatives in Global Perspective*. The MIT Press Cambridge, MA.
Yen, S. H., Lim, H. E., & Campbell, J. K. (2015). Age and productivity of academics: A case study of a public university in Malaysia. *Malaysian Journal of Economic Studies, 52*(1), 97–116.

Appendices

Appendix 1. Coefficients estimates.

	Quantitative productivity (number of papers)			*Qualitative productivity (citations of papers)*		
	WoS	*Scopus*	*RSCI*	*WoS*	*Scopus*	*RSCI*
percent of young	0.001	0.012***	0.013***	-0.029*	-0.016	0.012***
	(0.004)	(0.003)	(0.002)	(0.015)	(0.014)	(0.004)
percent of old	-0.0001	0.007**	-0.014***	-0.034**	-0.022*	-0.012***
	(0.004)	(0.003)	(0.002)	(0.014)	(0.013)	(0.004)
percent of young * leading status	0.026*	0.018	-0.011	-0.017	-0.007	0.008
	(0.015)	(0.013)	(0.007)	(0.027)	(0.024)	(0.012)
percent of old * leading status	0.047***	0.037***	0.010	0.068***	0.071***	0.063***
	(0.013)	(0.011)	(0.006)	(0.023)	(0.021)	(0.011)
leading status	-0.750	-0.347	0.116	0.301	-0.103	-1.166**
	(0.657)	(0.550)	(0.336)	(1.165)	(1.030)	(0.555)
offline students	0.032***	0.035***	0.002	0.026	0.033*	0.003
	(0.006)	(0.005)	(0.004)	(0.018)	(0.019)	(0.006)
students per faculty	-0.003	0.007*	0.013***	-0.002	0.008	0.020***
	(0.005)	(0.004)	(0.003)	(0.020)	(0.013)	(0.005)
students per faculty * leading status	0.007	-0.002	-0.007	0.006	0.004	-0.033**
	(0.018)	(0.016)	(0.010)	(0.037)	(0.029)	(0.016)
share of foreign	0.118***	0.088***	0.027**	0.227***	0.159***	0.031
	(0.023)	(0.017)	(0.012)	(0.058)	(0.041)	(0.025)
PhD students	0.018***	0.027***	-0.006*	0.063***	0.053***	0.031***
	(0.005)	(0.004)	(0.003)	(0.018)	(0.014)	(0.006)
hard profile	0.363***	0.500***	-0.093***	1.032***	1.020***	-0.040
	(0.073)	(0.054)	(0.034)	(0.302)	(0.264)	(0.068)
medicine profile	-0.225***	-0.087	-0.265***	-0.178	0.028	-0.555***
	(0.084)	(0.069)	(0.046)	(0.334)	(0.336)	(0.081)
agriculture profile	-0.111	-0.085	0.396***	-0.199	-0.191	0.443***
	(0.082)	(0.065)	(0.051)	(0.220)	(0.193)	(0.102)
art & sport profile	-1.004***	-0.789***	-0.326***	-0.967***	-0.653**	0.015
	(0.163)	(0.108)	(0.070)	(0.330)	(0.299)	(0.126)
year 2016	0.382***	0.383***	0.314***	0.804	0.805*	0.576***
	(0.105)	(0.074)	(0.041)	(0.546)	(0.440)	(0.086)

year 2017	0.903***	0.556***	0.457***	0.941***	0.779***	0.824***
	(0.089)	(0.066)	(0.042)	(0.336)	(0.271)	(0.088)
year 2018	0.993***	0.850***	0.490***	0.980***	1.082***	0.884***
	(0.088)	(0.064)	(0.043)	(0.316)	(0.269)	(0.081)
year 2019	1.142***	1.056***	0.631***	1.229***	1.338***	1.020***
	(0.082)	(0.063)	(0.045)	(0.299)	(0.256)	(0.084)
year 2020	1.137***	1.202***	0.704***	1.419***	1.547***	1.139***
	(0.086)	(0.062)	(0.045)	(0.291)	(0.248)	(0.086)
Constant	1.395***	1.152***	4.829***	4.229***	3.696***	5.617***
	(0.238)	(0.173)	(0.107)	(0.823)	(0.792)	(0.223)
Observations	2,164	2,164	2,164	2,164	2,164	2,164
Log-likelihood	-7,551.056	-8,228.448	-13,500.270	-11,712.490	-12,377.980	-17,231.650
Theta	1.793*** s.e.=0.063	2.387*** s.e.=0.083	3.306*** s.e.=0.098	0.618*** s.e.=0.017	0.768*** s.e.=0.021	1.090*** s.e.=0.030
AIC	15,142.110	16,496.900	27,040.540	23,464.970	24,795.960	34,503.290

Note: Bootstrapped standard errors in parentheses, if other not specified
Stars correspond to p-values: *p<0.1, **p<0.05, ***p<0.01

Appendix 2. Incidence rates estimates.

	Quantitative productivity (number of papers)			*Qualitative productivity (citations of papers)*		
	WoS	*Scopus*	*RSCI*	*WoS*	*Scopus*	*RSCI*
percent of young	1.001	1.012***	1.013***	0.972*	0.984	1.012***
	(0.853)	(0.0002)	(0.000)	(0.060)	(0.256)	(0.007)
percent of old	1.000	1.007**	0.986***	0.966**	0.979*	0.988***
	(0.987)	(0.027)	(0.000)	(0.012)	(0.100)	(0.003)
percent of young * leading status	1.026*	1.018	0.989	0.983	0.993	1.008
	(0.082)	(0.145)	(0.111)	(0.527)	(0.772)	(0.489)
percent of old * leading status	1.048***	1.038***	1.010	1.071***	1.073***	1.065***
	(0.001)	(0.001)	(0.102)	(0.004)	(0.001)	(0.000)
leading status	0.472	0.707	1.123	1.351	0.902	0.312**
	(0.254)	(0.529)	(0.730)	(0.797)	(0.921)	(0.036)
offline students	1.032***	1.036***	1.002	1.026	1.034*	1.003
	(0.000)	(0.000)	(0.545)	(0.151)	(0.076)	(0.628)
students per faculty	0.997	1.007*	1.013***	0.998	1.008	1.020***
	(0.489)	(0.067)	(0.000)	(0.913)	(0.521)	(0.000)
	1.007	0.998	0.993	1.006	1.004	0.968**

students per faculty * leading status	(0.682)	(0.904)	(0.509)	(0.876)	(0.878)	(0.040)
share of foreign	1.125***	1.092***	1.028**	1.255***	1.172***	1.032
	(0.000)	(0.000)	(0.020)	(0.000)	(0.000)	(0.208)
PhD students	1.018***	1.028***	0.994*	1.065***	1.054***	1.031***
	(0.001)	(0.000)	(0.077)	(0.001)	(0.000)	(0.000)
hard profile	1.438***	1.649***	0.912***	2.807***	2.772***	0.961
	(0.000)	(0.000)	(0.006)	(0.001)	(0.000)	(0.561)
medicine profile	0.799***	0.916	0.768***	0.837	1.028	0.574***
	(0.008)	(0.208)	(0.000)	(0.594)	(0.935)	(0.000)
agriculture profile	0.895	0.918	1.486***	0.820	0.826	1.557***
	(0.177)	(0.191)	(0.000)	(0.367)	(0.323)	(0.000)
art & sport profile	0.366***	0.454***	0.722***	0.380***	0.520**	1.015
	(0.000)	(0.000)	(0.000)	(0.004)	(0.029)	(0.906)
year 2016	1.466***	1.467***	1.368***	2.235	2.237*	1.779***
	(0.000)	(0.000)	(0.000)	(0.141)	(0.068)	(0.000)
year 2017	2.468***	1.744***	1.580***	2.562***	2.180***	2.280***
	(0.000)	(0.000)	(0.000)	(0.006)	(0.005)	(0.000)
year 2018	2.701***	2.340***	1.632***	2.665***	2.950***	2.421***
	(0.000)	(0.000)	(0.000)	(0.002)	(0.000)	(0.000)
year 2019	3.134***	2.876***	1.880***	3.418***	3.810***	2.773***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
year 2020	3.119***	3.328***	2.022***	4.131***	4.695***	3.124***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	4.034***	3.165***	125.048***	68.630***	40.284***	275.110***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	2,164	2,164	2,164	2,164	2,164	2,164
Log-likelihood	-7,551.056	-8,228.448	-13,500.270	-11,712.490	-12,377.980	-17,231.650
Theta	1.793*** s.e.=0.063	2.387*** s.e.=0.083	3.306*** s.e.=0.098	0.618*** s.e.=0.017	0.768*** s.e.=0.021	1.090*** s.e.=0.030
AIC	15,142.110	16,496.900	27,040.540	23,464.970	24,795.960	34,503.290

Note: P-values in parentheses, if other not specified
Stars correspond to p-values: *p<0.1, **p<0.05, ***p<0.01
Bootstrapped standard errors were used for p-values calculation

https://doi.org/10.51408/issi2025_072

AI on AI: Exploring the Utility of GPT as an Expert Annotator of AI Publications

Autumn Toney-Wails[1], Christian Schoeberl[2], James Dunham[3]

[1]*autumn.toney@georgetown.edu*
Georgetown University, Washington D.C. (USA)
SciTech Strategies, Inc. (USA)

[2]*christian.schoeberl@georgetown.edu*
Georgetown University, Washington D.C. (USA)

[3] *james.dunham@georgetown.edu*
Georgetown University, Washington D.C. (USA)

Abstract

Identifying scientific publications that are within a dynamic field of research often requires costly annotation by subject-matter experts. Resources like widely-accepted classification criteria or field taxonomies are unavailable for a domain like artificial intelligence (AI), which spans emerging topics and technologies. We address these challenges by inferring a functional definition of AI research from existing expert labels. We then evaluate state-of-the-art chatbot models on the task of expert data annotation. Using the arXiv publication database as ground truth, we experiment with prompt engineering for OpenAI's GPT chatbot models (3.5-Turbo and 4) to identify an alternative, automated expert annotation pipeline that assigns AI labels with 94% accuracy. For comparison, we fine-tune SPECTER, a transformer language model pre-trained on scientific publications, using arXiv publications that achieves 96% accuracy (only 2% higher than GPT-4) on classifying AI publications. Our results indicate that with effective prompt engineering, chatbots can be used as reliable data annotators even where subject-area expertise is required. To evaluate the utility of chatbot-annotated datasets on downstream classification tasks, we fine-tune a classifier on GPT-labeled data that outperforms the classifier fine-tuned on arXiv data by nine percentage points, achieving 82% accuracy.

Introduction

Analyzing scholarly literature provides insight into important features of a research field: identifying active research communities, tracking recent advances or breakthroughs, and mapping the translation of basic research into applications. A significant challenge for field-level analyses is the lack of clearly-defined, widely-accepted criteria to identify relevant scientific text, particularly when a field contains rapidly emerging topics and technologies (Kurzweil, 1985; Suominen and Newman, 2017). Artificial intelligence (AI) is one such research field. The challenge of identifying AI research is not new; as described by Schank (1987), "Because of the massive, often quite unintelligible publicity that it gets, artificial intelligence is almost completely misunderstood by individuals outside the field. Even AI's practitioners are somewhat confused about what AI really is."

Currently, almost 40 years after Schank questioned what AI was, identifying AI research is still an ambiguous task. Definitions of AI vary from academia, industry, and government, creating a challenge for researchers and policymakers when trying

to conduct bibliometric studies, forecast technological capabilities, evaluate global leadership, or develop effective policy for AI systems and models (Kurzwell, 1985; Toney and Dunham 2022; Grace et al., 2018; Krafft et al., 2020; Cave and ÓhÉigeartaigh, 2018). While establishing a succinct definition or granular taxonomy of AI as a research field is outside the scope of this paper, we present an approach to automatically identify and classify AI research that leverages state-of-the-art large language models as expert annotators. We propose a generalizable framework for classification tasks that do not have a clearly defined labelling convention, and thus, are not amenable to costly, error-prone manual annotation.

We derive an AI definition from published research activity over the past decade. Using a collection of author-identified AI research publications from the open-access arXiv database, we create a subset of AI-related scientific research publications (**AI-arXiv**) as a ground-truth labeled dataset. Our approach uses what subject-area experts have identified as relevant to AI research, reducing bias. The labels are assigned by the authors and thus reflect the evolution of the field; they are not bound by a static or dated definition.

With scientific publications' titles and abstracts as input text, we use two transformer language models pre-trained on scientific text, SciBERT (Beltgay et al., 2019) and SPECTER (Cohan et al., 2020) for AI publication classification. For this classification task we use two scholarly literature datasets: (1) arXiv,[1] as it contains author-assigned subject categories and (2) OpenAlex (Priem et al., 2022), as it contains the majority of scholarly literature but requires expert annotation. We establish an AI classification accuracy baseline using AI-arXiv to fine-tune both language models. Then, we explore the utility of OpenAI's GPT models as an automated expert annotator of AI publications using a zero-shot annotation prompt. We design a series of prompts with personas of varying levels of expertise (reader, researcher, and subject-matter expert) for the AI publication annotation task. We experiment with both GPT-3.5-Turbo and GPT-4 due to the significant cost difference.

We compare the GPT chatbot models' labeling accuracies and model performances to our baseline AI-arXiv classifier and evaluate how accurately the GPT models assign labels to AI-related arXiv publications. Selecting the most reliable and accurate zero-shot prompt and GPT model, we generate a dataset of GPT-labeled publications (**AI-GPT**) to train a new publication classifier and compare it to the baseline AI-arXiv classifier. Our results show that both GPT-3.5-Turbo and GPT-4 achieve 94% accuracy in data labeling, compared to the baseline AI-arXiv classifier which achieves 96% accuracy, suggesting that chatbots can be effectively used as expert annotators with reliable results. Evaluating the AI-arXiv and AI-GPT classifiers on a set of publications from the top 13 AI conferences, we find that the AI-GPT classifier outperforms the AI-arXiv classifier by nine percentage points, achieving 82% accuracy.

We summarize our contributions as follows: 1) we design an experimental framework to evaluate a chatbot's utility as an expert annotator, 2) we propose and

[1] https://arxiv.org

discuss the merits of using a crowd-sourced AI definition derived from experts, and 3) we evaluate our framework on an AI research publication classification task.

Background and Motivation

Establishing a field definition or taxonomy is challenging for areas of research that encompass rapidly emerging topics and technologies, with AI being no exception. Since the term AI was first coined by McCarthy in 1955, researchers have addressed the question—what *is* AI?—by surveying existing research and proposing frameworks and definitions (Schank, 1987; Kurzweil, 1985; Russell, 2010; Fast and Horvitz, 2017; Martınez-Plumed et al., 2018; Krafft et al., 2020; Shukla et al., 2019). We highlight two notable instances of computer scientists working to intentionally think through this question and answer it, both arriving at similar conclusions. Schank (1987) responded to this question pragmatically, stating that the definition of AI to a given researcher or organization depends directly on their research goals and methodology to design and implement their AI model. Kurzweil (1985) acknowledged that academia and industry are at odds with each other in forming a consensus about what AI is, correctly assessing that resolving this controversy is not likely in the near future.

Although a clear, widely-accepted definition of AI has not been established, there are features of AI as a field that are agreed-upon, namely its direct relationship to machine learning (ML). Many publications have distinguished the two fields by describing ML as an application of AI (Alpaydin 2016; Gröger, 2021; Woolridge, 2022). In this work, we incorporate this notion, while we do not use the two terms (AI and ML) as synonyms, we consider ML to be a majority subset of AI research.

Defining and Identifying AI Research

Prior research has attempted to establish methodologies for identifying AI based on varying criteria. Goa et al. (2021) generate a list of 127 AI journals for publication analysis. In a similar approach, Martinez et al. (2018) study publications from AAAI and IJCAI conferences and Shukla et al. (2019) study publications from the Engineering Applications of Artificial Intelligence. Using query-based methods, Niu et al. (2016) selects publications containing the term “artificial intelligence” and Miyazaki and Ryusuke (2018) develop a query with 43 search terms (e.g., “machine learning” and “facial recognition”).

Scientific publication databases often provide research subject areas that are not assigned manually by experts. One common example is the use of Microsoft Academic Graph's (MAG) fields of study, a hierarchy of research topics organized into five levels of granularity (level 0 - 4) (Sinha et al., 2015; Shen et al., 2018). These topic assignments allow for bibliometric analysis on research topics across all of science, but the topic assignment is unsupervised and based on embedding similarity. While this mitigates annotator bias, it relies on a static definition of AI derived from Wikipedia descriptions of concepts. Additionally, these assignments are not evaluated against ground-truth data.

Identifying AI research based on a set of publication venues or keywords restricts analysis to venues and terms that were relevant to the field at the time the study was conducted. These methods are also at risk of creating a narrowly scoped set of papers, ignoring more general or cross-disciplinary research that is relevant to the field. We address these shortcomings by selecting publications with author-assigned research categories, where the categories are assigned at the time of publication. This approach uses what subject-area experts consider to be relevant to AI and ML, incorporating how the field and its activity has evolved.

Large Language Models as Expert Annotators

With the ability to perform natural language processing tasks with human-like reasoning, chatbots have enabled users to interact with generative AI systems that can produce human-like responses. However, despite chatbots' potential to respond with correct and reliable results, they are prone to respond with *hallucinations*, a term representing seemingly random, non-factual or incoherent chatbot responses (Dziri et al., 2022; Ji et al., 2023). These hallucinations resulted in research focused on how to effectively leverage a chatbot as a reliable data annotator via prompt engineering, as chatbots can be cost effective against human annotators, but concerningly less reliable.

Wang et al. (2021) find GPT-3 to be on average 10 times cheaper than human annotators when compared to Google Cloud Platform prices. Gilardi et al. (2023) evaluated ChatGPT on over 6,000 tweets and news articles for numerous classification tasks, including relevance, topic assignment, and stance detection. Using six trained annotators, they established ground-truth labels for comparison to Mechanical Turkers and ChatGPT and found that ChatGPT achieved higher annotation performance and was an estimated 30 times cheaper than manual annotators.

Kim et al. (2023) evaluate ChatGPT's ability to label the strength of a claim as causal, conditional causal, correlational, or no relationship. The authors found that ChatGPT did not achieve state-of-the-art classification performance, concluding that chatbots have promising annotation capacity but improvement is needed for causal scientific reasoning. Our work analyzes chatbots' abilities when leveraged as a reliable, expert annotator on a zero-shot task. We consider our annotation task to be straight-forward, requiring less reasoning than evaluating causal scientific claims, but more expertise than a typical human annotator might have. Our annotation experiments investigate optimal prompt design for similar annotation tasks, presenting a format that can be adapted to other domains.

Experimental Design

We design a framework to evaluate the utility of chatbots as expert annotators through prompt engineering and classification model performance, as shown in Figure 1. Each step in this process can be customized to a specific classification task, provided that there is a baseline labeled dataset or a comparable evaluation task for prompt engineering. In a small data task, chatbot annotation can replace the classifier

model. However, in this work we focus on tasks that use large datasets, which would be time-intensive and costly for a chatbot to annotate entirely.

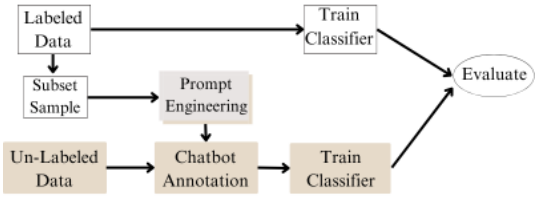


Figure 1. Chatbot annotation experimental framework diagram.

Scientific Publication Classifier

We experiment with two publicly available transformer language models: SciBERT and SPECTER. SciBERT is a pre-trained language model based on the Bidirectional Encoder Representations from Transformers (BERT) model and trained on a sample of 1.14M papers from Semantic Scholar (using full-text) (Beltagy et al., 2019). The Scientific Paper Embeddings using Citation-informed Transformers (SPECTER) improves on SciBERT by incorporating the citation graph that exists between academic publications (Cohan et al., 2020). Compared to SciBERT, SPECTER decreases training times while maintaining (or exceeding) performance, particularly for classification tasks.

We fine-tune both pre-trained scientific publication classifiers in the same way, changing transformer models and datasets for each experiment but maintain training, testing, and validation datasets (details on building the classifiers can be found on our GitHub repository[2]). We use train, test, and validation dataset splits of 70%, 15%, and 15% respectively.

Data Annotation and Prompt Engineering

To automatically identify AI research publications we leverage state-of-the-art LLMs' chatbot feature to assign binary (AI or non-AI) labels given a publication's title and abstract. This requires experimentation in prompt engineering to select a prompt that will produce reliable labels that are usable on their own (e.g., treating the chatbot as the classifier) or functional for generating training data (e.g., creating a labeled dataset). We aim to identify a single, zero-shot prompt that will produce accurate and parsable responses to initiate an automated annotation pipeline.

We use the GPT-3.5-Turbo and GPT-4 chatbots, as GPT-4 is a more robust model but is approximately 20 times more expensive to query than GPT-3.5-Turbo. Using the *openai* Python package, we run our prompt engineering experiments with *temperature* set to 0, which indicates that the most likely output from the model should be selected. We design a series of prompts that provide increasing specificity of the AI expertise we ask the GPT chatbot models to personify: a reader, a researcher, and a subject-matter expert. Table 1 lists the nine prompts that we test. Each of the three personas are tested with variations to the instructions, providing

[2] https://github.com/georgetown-cset/arxiv-classifiers

the chatbot with awareness of non-relevant publications and clarity on how to annotate. These prompts result in consistent and parsable responses, enabling us to compare their performance across chatbot models and prompts.

Table 1. Zero-shot chatbot prompt variations for reliable data annotation experiments.

Prompt Type	*Prompt*
Reader, Researcher, Expert	You are a **[persona type]** in AI/ML, and you are given an annotation task. Given a publication's title and abstract, assign a AI or Non-AI label determining if the publication belongs to the field of AI/ML research and a predicted probability of relevance. Provide just the label and prediction in your answer.
Uncertainty	You are a **[persona type]** in AI/ML, and you are given an annotation task. Given a publication's title and abstract, assign a AI or Non-AI label determining if the publication belongs to the field of AI/ML research and a predicted probability of relevance. **Some papers may be in STEM fields but not exactly AI, please assign AI only if you are confident**. Provide just the label and probability in your answer
Uncertainty and Clarity	You are a **[persona type]** in AI/ML, and you are given an annotation task. Based on the title and abstract of an academic publication, assign a label "AI" or "Non-AI" indicating whether the publication belongs to the field of AI/ML research. Also assign a score between 0 and 1 that describes how confident you are in the label. **Some papers may be in STEM fields but not exactly in AI/ML. Please assign the "AI" label only if you are confident. Otherwise, assign the "Non-AI" label and quantify your uncertainty in the score.** Respond only with the label and the score.

Classifier Performance Evaluation

We include an evaluation task that is separate from comparing the validation performance of the two classification models. This evaluation task should compute model performance on a new dataset to evaluate the generalizability of the models, as one goal of using chatbots for data annotation is generating a more representative training dataset. For our domain application, we follow Toney and Dunham (2022) and use a set of publications that appeared in one of the 13 top AI and ML conferences from CSRankings:[3] 1) AAAI Conference on Artificial Intelligence, 2) International Joint Conference on Artificial Intelligence, 3) IEEE Conference on Computer Vision and Pattern Recognition, 4) European Conference on Computer

[3] https://csrankings.org/

Vision, 5) IEEE International Conference on Computer Vision, 6) International Conference on Machine Learning, 7) International Conference on Knowledge Discovery and Data Mining, 8) Neural Information Processing Systems, 9) Annual Meeting of the Association for Computational Linguistics, 10) North American Chapter of the Association for Computational Linguistics, 11) Conference on Empirical Methods in Natural Language Processing, 12) International Conference on Research and Development in Information Retrieval, and 13) International Conference on World Wide Web.

Scholarly Literature Datasets

We use two open-source scholarly literature datasets in our experiments: arXiv and OpenAlex. Here we describe the details of each data source, as well as define how we generate two AI publication datasets for experimentation: **AI-arXiv** and **AI-OpenAlex**.

arXiv Dataset

Hosting over 2 million scientific publications across 8 research fields (computer science, economics, electrical engineering and systems science, mathematics, physics, quantitative biology, quantitative finance, and statistics), arXiv is a useful resource for classification tasks on scientific text, as all publications are assigned research area categories by authors. arXiv's Computing Research Repository (CoRR) lists 39 sub-categories including artificial intelligence and machine learning. CoRR editors review each publication and its assigned research categories, so we consider these labels to be expert-annotated.

We generate the AI-arXiv dataset, comprising publications since 2010 that CoRR identifies as being AI-related using their author-assigned research categories.[4] We include a publication in the AI-arXiv dataset if it was labeled with at least one of the following research topics: Artificial Intelligence (cs.AI), Computation and Language (cs.CL), Computer Vision and Pattern Recognition (cs.CV), Machine Learning (cs.LG, stat.ML), Multiagent Systems (cs.MA), and Robotics (cs.RO). For our binary classification task (AI vs. non-AI), we use publications that do not contain one of the seven AI-related labels as non-AI publications.

The dataset begins in 2010, but the majority of AI-related publications (80%) are from 2018 and later, as shown in Figure 2. This graph illustrates the rapid influx of AI-related publications and highlights that AI-arXiv mainly represents research activity from the past five years.

[4] https://arxiv.org/category_taxonomy

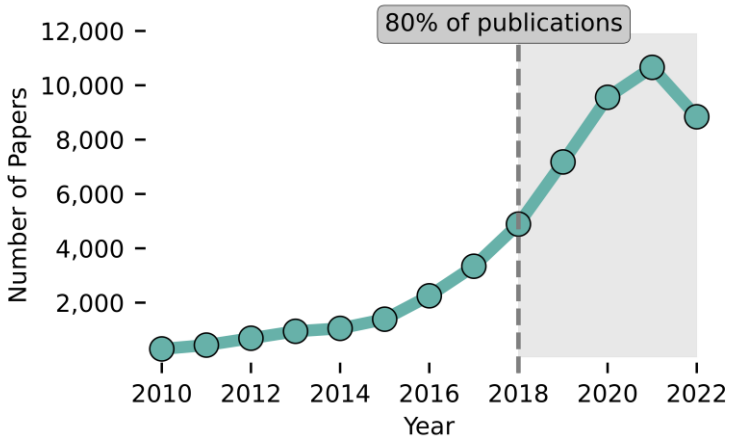


Figure 2. Number of AI-arXiv by publication year. Data accessed on 10-13-2022, thus 2022 is incomplete.

Authors can assign a primary subject category and cross-post under additional categories (i.e., a publication may be assigned more than one category). Table 2 displays the number of publications by their primary arXiv research category and by cross-post research categories. The top two most frequent categories are machine learning (cs.LG) with 120,525 publications and computer vision (cs.CV) with 82,760 publications. Multiagent systems (cs.MA) is the least frequent category with 5,304 publications.

Table 2. Number of publications by primary and cross-post research category type.

Post Type	*AI*	*CL*	*CV*	*LG*	*MA*	*RO*	*ML*
Primary	14,615	28,914	63,016	55,991	1,801	14,246	13,516
Cross-Post	36,332	7,860	19,744	64,534	3,503	6,381	42,412

OpenAlex Dataset

Containing over 240 million scientific publications across all fields of science, we use OpenAlex as our un-labeled publication dataset that we sample from for our chatbot annotation task. To maintain consistency with the AI-arXiv dataset we restrict publication year to 2010 or later and to refine the OpenAlex publications that we use for fine-tuning, we require at least one citation per publication, resulting in 114,635,253 publications. OpenAlex provides concepts[5] which are structured similarly to Microsoft Academic Graph's field of study taxonomy that are automatically assigned to each publication using document embedding similarity; category assignments are not validated against ground truth data for all field categories, thus the assignments contain noise. There are 19 concepts at the most general level (e.g., physics and computer science) and there are 283 subtopics at the next level of granularity (e.g., AI and ML). Because these concepts are commonly

[5] https://docs.openalex.org/api-entities/concepts

used for labels in classification tasks-—and specifically used by Cohen et al. (2020) evaluate SPECTER—we generate a set of AI publications to compare against the AI-arXiv dataset. We select concepts that directly map to the arXiv categories; publications with artificial intelligence, machine learning, computer vision, and natural language processing listed as the top field compose the AI-OpenAlex dataset, totaling 4,305 AI publications.

Results and Evaluation

We present our results across all experiments by organizing the following sections to describe our selection of the final AI publication classifier, evaluation of GPT as a data annotator of AI publications, and an overall comparison of fine-tuned SPECTER classifiers on the top 13 AI conferences dataset.

AI-arXiv Classifier

Using the AI-arXiv dataset we fine-tune the SciBERT and SPECTER language models on a binary classification task (AI or non-AI) given a publication's title and abstract. We run this initial experiment on a 10% sample of the AI-arXiv dataset (over 150K publications) fine-tuning both SciBERT and SPECTER. We only fine-tune SciBERT on the AI-arXiv dataset, as Cohan et al. (2020) showed that SPECTER (achieving 80%) outperformed SciBERT (achieving 72%) on classifying publications by MAG's most general fields of study. Table 3 displays the three classifier performance metrics.

Table 3. 10% sample of the AI-arXiv dataset and the AI-OpenAlex dataset: fine-tuned model performances with SciBERT and SPECTER.

Model	*Accuracy*	*Precision*	*Recall*	*F1*
$\text{SciBERT}_{\text{arXiv}}$	0.96	0.88	0.86	0.87
$\text{SPECTER}_{\text{MAG}}$	0.93	0.93	0.60	0.73
$\text{SPECTER}_{\text{arXiv}}$	0.96	0.87	0.88	0.88

Table 3 shows that SPECTER fine-tuned on AI-arXiv outperforms the other two models in all performance metrics, with approximately the same training time and cost as SciBERT, thus we select SPECTER as our transformer language model for classification tasks. Table 4 displays the total counts for the train, test, and validation splits for the full AI-arXiv dataset. The SPECTER model fine-tuned on the full AI-arXiv dataset achieves similar results (shown in Table 8) to the 10% sample model in Table 3.

Table 4. Train, test, and validation dataset split counts for the AI-arXiv dataset, by label.

Split	*Total Count*	*AI*	*Non-AI*
Train	1,093,647	173,292	920,355
Test	234,353	37,136	197,217
Validation	234,353	37,042	197,311

GPT for Data Annotation

With the nine prompt variations listed in Table 1, we experiment with GPT-3.5-Turbo and GPT-4 as expert data annotators. We sample 5,000 publications from the AI-arXiv dataset, and due to cost we only query GPT-4 with 2,500 publications for our initial chatbot comparison. In all chatbot responses, we are able to parse the response into a binary label (AI/non-AI) and a predicted relevance probability (with the values always ranging between 0 and 1).

The system prompt significantly impacts a chatbot's ability to accurately annotate publications, as shown in Table 5. GPT-4 provides consistent performance across all prompt variations; however, the best GPT-3.5-Turbo prompts are able to produce the same accuracy results. We find that GPT-3.5-Turbo has a stronger improvement when including language surrounding uncertainty and clarity in the prompt. Adding additional prompt language regarding the annotation task increases accuracy by 13.8 percentage points on average; however, the persona shifts have minimal effect across both chatbot models and prompt variation. In contrast to GPT-3.5-Turbo, GPT-4 has significantly less improvement when including the uncertainty and clarity clauses.

Table 5. GPT chatbot model comparison across all nine prompts for data annotation. – denotes the baseline prompt with no additional clauses, +U denotes the prompt including uncertainty, and +UC denotes the prompt including uncertainty and clarity.

GPT Model	*Accuracy*								
	Reader			*Researcher*			*Expert*		
	-	*+U*	*+UC*	-	*+U.*	*+UC*	-	*+U*	*+UC*
3.5-Turbo	.79	.91	.92	.76	.91	.92	.78.	.91	.90
4	.91	.94	.92	.91	.92	.92	.91.	.94	.94

We select the expert with uncertainty and clarity prompt for further experimentation. To investigate if there is a subject area difference between GPT-3.5-Turbo's and GPT-4's performances, we compared the annotation accuracy of the best prompt. Table 6 displays the accuracies by arXiv research category across both GPT models. GPT-3.5-Turbo has the highest annotation accuracies on machine learning (89%), NLP (89%), and computer vision (87%), with multiagent systems (67%) and robotics (75%) categories having the worst performance. Similarly, GPT-4 performs the

worst on multiagent systems (89%) and robotics (84%), but performs the best on machine learning (99%) and AI (98%).

Table 6. Label accuracies by GPT model and arXiv categories, including primary and cross-posted categories (publications can be counted in multiple categories).

GPT Model	*Accuracy*								
	AI	*CL*	*CV*	*LG*	*MA*	*RO*	*ML*	*None*	*Overall*
3.5-Turbo	.84	.89	.87	.89	.67	.75	.81	.93	.92
4	.98	.93	.93	.99	.89	.84	.97	.94	.94

Lastly, we evaluate the predicted probabilities of relevance that the chatbots provided. All responses included a probability value, with all values being correctly bounded by 0 and 1. The majority of values were either 0.95 or 0.2, thus we present the median predicted probably by GPT model and predicted class in Figure 3. We expect low median values for false negatives (FN) and true negatives (TN), indicating that the chatbot interpreted the publication's title and abstract as being non-relevant to AI research. We find that GPT-3.5-Turbo has expected results, with the negative class having low predicted probability (0.2 for FN and TN) and the positive class having high predicted probabilities (0.95 for TP and 0.9 for FP). In contrast, GPT-4 has the highest predicted probabilities for true positive (0.95), true negative (0.95), and false positive (0.85), with only a low predicted probability for false negatives (0.2).

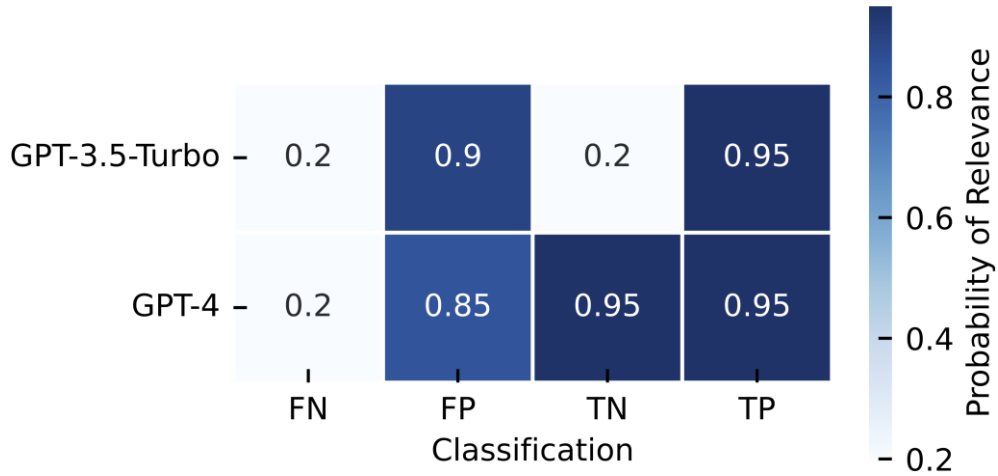


Figure 3. Median predicted probability of relevance by GPT model across classification types.

Using a 76,000-publication random sample of OpenAlex, we prompt GPT-4 with the expert with uncertainty and clarity prompt to generate **AI-GPT**, a GPT annotated AI set of publications.

AI-GPT Classifier

We fine-tune the SPECTER model with the AI-GPT dataset to compare results with the AI-arXiv fine-tuned model. Table 7 displays the train, test, and validation AI-GPT dataset splits, with a significant class imbalance towards non-AI papers. However, this is consistent with the representation of peer-reviewed AI publications in the context of all of science, with AI publications representing approximately 3.8% of all scientific literature (Zhang et al., 2021).

Table 7. Train, test, and validation dataset split counts for the AI-GPT dataset by labels.

Split	*Total Count*	*AI*	*Non-AI*
Train	53,459	1,288	52,171
Test/Validation	11,456	276	11,180

Table 8 displays the comparison between the SPECTER model fine-tuned on the AI-arXiv and AI-GPT datasets. We find that fine-tuning SPECTER using AI publications that were automatically annotated by GPT-4 produces an overall accuracy that is four percentage points lower than the expert labeled data from arXiv, with the F1 score being 8 percentage points lower.

Table 8. Fine-tuned SPECTER results using the AI-arXiv and AI-GPT datasets for AI classification.

Model	*Accuracy*	*Precision*	*Recall*	*F1*
$SPECTER_{arXiv}$	0.96	0.89	0.87	0.88
$SPECTER_{GPT\text{-}4}$	0.92	0.70	0.92	0.80

Classifier Evaluation

To evaluate the utility of chatbots as expert annotators and AI-related arXiv publications as a functional definition of AI research, we compare the AI-GPT, AI-arXiv, and AI-OpenAlex fine-tuned classifiers on a new dataset that contains publications from the top 13 AI conferences. Table 9 presents the models' accuracies by conference venue as well as the overall accuracy considering all AI conference papers as a set. We find that the AI-GPT model outperforms the AI-OpenAlex and AI-arXiv model across all venues, producing an overall accuracy of 82%. While accuracy by conference varies significantly across the classifiers, we find that CVPR, EMNLP, and ICCV have the highest accuracies for all three classifiers and that WWW and SIGIR have the lowest. None of the classifiers perform the best on AAAI or IJCAI, which are the two most explicitly AI-related conferences.

Table 9. AI-OpenAlex, AI-arXiv, and AI-GPT classification accuracies by conference venue.

Venue	*Num. of Papers*	AI-OpenAlex *Accuracy*	*AI-arXiv Accuracy*	*AI-GPT Accuracy*
NeurIPS	10,999	0.45	0.73	0.84
AAAI	10,446	0.53	0.88	0.89
IJCAI	9,700	0.32	0.68	0.70
ICML	6,192	0.46	0.70	0.82
CVPR	3,381	0.89	0.86	0.95
SIGIR	2,492	0.16	0.21	0.59
NAACL	1,834	0.62	0.86	0.93
ICCV	1,619	0.92	0.90	0.99
WWW	1,241	0.12	0.21	0.45
SIGKDD	1,064	0.31	0.60	0.84
ACL	932	0.71	0.86	0.93
EMNLP	839	0.79	0.93	0.95
ECCV	64	0.67	0.28	0.72
Overall	50,803	0.48	0.73	0.82

Discussion

Defining Research Fields via Expert Crowd-sourcing

A known limitation in text classification for AI research is the key step of identifying a set of labeled publications for classifier training. Designing a manual annotation task with either few expert or many non-expert annotators places the responsibility of defining what AI is on the authors, as they need to develop annotation instructions for the labeling task. Implementing unsupervised natural language processing techniques, such as topic modeling or document embedding clustering, lacks the transparency and reproducibility that can be achieved with a supervised classification model using reliably labeled data.

Our approach treats the authors as experts in the field and considers the publications and potential evolution of how authors assign labels as a time-relevant representation of research activity, opposed to relying on a static or narrow field definition. We find that our ground-truth data (AI-arXiv) is functional as training data for a generalized AI classification model in comparison to the AI-OpenAlex dataset, which labels publications by document embedding similarity. The AI-arXiv fine-tuned model is restricted to negative (non-AI) samples that are in STEM fields, meaning that the separation between relevant and non-relevant AI research is scoped to publications that would be more likely to blur the field's boundary lines. In comparison, the AI-OpenAlex and AI-GPT datasets have negative samples that span all of science, which could prove more optimal for classifier generalizability.

Evaluating the Utility of Chatbots as Annotators

A challenge with chatbots is their inconsistency and tendency to hallucinate. These two flaws are of particular issue in an annotation task, where reliable responses and

consistent reasoning are necessary. An additional challenge when working with chatbots is the lack of transparency in the LLM's training data. For example, some tasks might prove to be more suitable to chatbots as annotators because the underlying LLM was trained on large amounts of relevant data, or in our annotation task, the same data we are asking the GPT model to label.

While we do not explore in depth methods to uncover what scientific publications the GPT models might have observed during training, we do provide various ways of analyzing the responses from GPT. Our annotation results from prompt engineering indicate that while selecting the right persona is important for customized GPT performance, it is also necessary to design a prompt that encourages reasoning. We found that including specific uncertainty and clarity clauses in our prompts boosted GPT-3.5-Turbo's performance to be comparable with GPT-4, whereas the changes in expertise did not. Additionally, we explored including instructions for the chatbot to respond with a predicted probability of relevance for every title and abstract. While both GPT models consistently understood the task of responding with a probability, and always with a value between 0 and 1, GPT-3.5 responded more reliably. We found that GPT-4 responded with a median probability of 0.95 for labeling non-AI publications correctly.

Conclusion

In this work we investigate the utility of a chatbots as expert annotators by evaluating their annotation agreement with ground-truth data and their model performance on downstream classification tasks for identifying AI research publications. We address the challenge of identifying AI, a rapidly emerging research field with no clear definition, by leveraging expert, crowd-sourced data on arXiv. We find that GPT models are able to achieve high accuracy as expert annotators on AI publications, producing reliable and parsable responses necessary for an annotation task. Our prompt engineering experiments indicate that chatbots have the highest performance when the prompt includes a relevantly-scoped persona (e.g., AI researcher or subject-matter expert) as well as details on how to consider edge cases (e.g., language describing how to consider uncertainty or providing clarity on the annotation task). We also find that GPT models' labels can be reliably used in downstream classification tasks as training data. Our experiments show that even in large datasets with no underlying labels, GPT models can provide a functional boundary between positive and negative examples. Collectively, these findings signal the ability of chatbots to provide scalable and efficient data annotation for bibliometric analysis, upon which more complex models can be built.

References

Alpaydin, E. (2016). Machine learning: The new AI. MIT press.

Artificial intelligence (AI) vs. machine learning (ML). (n.d.). https://cloud.google.com/learn/artificial-intelligence-vs-machine-learning.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. EMNLP.

Cave, S., & ÓhÉigeartaigh, S. S. (2018). An AI race for strategic advantage: rhetoric and risks. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 36–40.

Clement, C. B., Bierbaum, M., O'Keeffe, K. P., & Alemi, A. A. (2019). On the use of arXiv as a dataset. arXiv preprint arXiv:1905.00075.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). SPECTER: document-level representation learning using citation-informed transformers. ACL.

Dziri, N., Milton, S., Yu, M., Zaiane, O., & Reddy, S. (2022). On the origin of hallucinations in conversational models: Is it the datasets or the models? Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 5271–5285.

Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. Proceedings of the AAAI Conference on Artificial Intelligence, 31.

Gao, F., Jia, X., Zhao, Z., Chen, C.-C., Xu, F., Geng, Z., & Song, X. (2021). Bibliometric analysis on tendency and topics of artificial intelligence over last decade. Microsystem Technologies, 27, 1545–1557.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. Proceedings of the National Academy of Sciences, 120(30), e2305016120.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. Journal of Artificial Intelligence Research, 62, 729–754.

Gröger, C. (2021). There is no AI without data. Communications of the ACM, 64(11), 98–108.

Huang, Y., Schuehle, J., Porter, A. L., & Youtie, J. (2015). A systematic method to create search strategies for emerging technologies based on the web of science: Illustrated for 'big data'. Scientometrics, 105, 2005–2022.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

Kim, Y., Guo, L., Yu, B., & Li, Y. (2023). Can ChatGPT understand causal language in science claims? Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, 379–389.

Krafft, P., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020). Defining AI in policy versus practice. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 72–78.

Kurzweil, R. (1985). What is artificial intelligence anyway? As the techniques of computing grow more sophisticated, machines are beginning to appear intelligent—but can they actually think? American Scientist, 73(3), 258–264.

Martínez-Plumed, F., Loe, B. S., Flach, P., ÓhÉigeartaigh, S., Vold, K., & Hernández-Orallo, J. (2018). The facets of artificial intelligence: A framework to track the evolution of AI.

Miyazaki, K., & Sato, R. (2018, August). Analyses of the technological accumulation over the 2nd and the 3rd AI boom and the issues related to AI adoption by firms. In 2018 Portland International Conference on Management of Engineering and Technology (PICMET) (pp. 1-7). IEEE.

Mogoutov, A., & Kahane, B. (2007). Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking. Research Policy, 36(6), 893–903.

Mustafa, G., Usman, M., Yu, L., Afzal, M. T., Sulaiman, M., & Shahid, A. (2021). Multi-label classification of research articles using word2vec and identification of similarity threshold. Scientific Reports, 11(1), 21900.

Niu, J., Tang, W., Xu, F., Zhou, X., & Song, Y. (2016). Global research on artificial intelligence from 1990–2014: Spatially-explicit bibliometric analysis. ISPRS International Journal of Geo-Information, 5(5), 66.

OpenAI. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv preprint arXiv:2205.01833.

Russell, S. J. (2010). Artificial intelligence: a modern approach. Pearson Education, Inc.

Sachini, E., Sioumalas-Christodoulou, K., Christopoulos, S., & Karampekios, N. (2022). AI for AI: Using AI methods for classifying AI science documents. Quantitative Science Studies, 1–14.

Schank, R. C. (1987). What is AI, anyway? AI Magazine, 8(4), 59–59.

Shen, Z., Ma, H., & Wang, K. (2018). A web-scale system for scientific knowledge exploration. Proceedings of ACL 2018, System Demonstrations, 87–92.

Shukla, A. K., Janmaijaya, M., Abraham, A., & Muhuri, P. K. (2019). Engineering applications of artificial intelligence: A bibliometric analysis of 30 years (1988–2018). Engineering Applications of Artificial Intelligence, 85, 517– 532.

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. (, & Wang, K. (2015). An overview of Microsoft Academic Service (MAS) and applications. Proceedings of the 24th International Conference on World Wide Web, 243– 246.

Suominen, A., & Newman, N. C. (2017). Exploring the fundamental conceptual units of technical emergence. 2017 Portland International Conference on Management of Engineering and Technology (PICMET), 1–5.

Sweeney, L. (n.d.). That's AI?: A history and critique of the field.

Toney, A., & Dunham, J. (2022). Multi-label classification of scientific research documents across domains and languages. Proceedings of the Third Workshop on Scholarly Document Processing, 105–114.

Wang, S., Liu, Y., Xu, Y., Zhu, C., & Zeng, M. (2021). Want to reduce labeling cost? GPT-3 can help. Findings of the Association for Computational Linguistics: EMNLP 2021, 4195–4205.

West, D. M., & Allen, J. R. (2018). How artificial intelligence is transforming the world. Report. April, 24, 2018.

Woolridge, M. (2022). A brief history of artificial intelligence: What it is, where we are, and where we are going. Flatiron Books.

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., Sellitto, M., et al. (2021). The AI index 2021 annual report. arXiv preprint arXiv:2103.06312.

https://doi.org/10.51408/issi2025_017

Almost Always Unequal: Co-Authors' Contributions to Scientific Publications

Paul Donner[1], Philippe Vincent-Lamarre[2]

[1]*donner@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW), Schützenstrasse 6a, 10117 Berlin (Germany)

[2]*Philippe.Vincent-Lamarre@cihr-irsc.gc.ca*
Canadian Institutes of Health Research (CIHR) 160 Elgin Street, K1A 0W9, Ottawa (Canada)

Abstract

Scientific work has become increasingly organized as teamwork and most research publications are now joint work of several co-authors. While of utmost importance for fair and valid research evaluation, the quantitative patterns of relative work contribution by team members to co-authored publications have remained opaque. Here we present an empirical study of contribution patterns. We analyze a large data set of author-provided percent contribution claims for co-authored scientific publications submitted as part of applications to scholarship programs. We find that the distribution of work input in co-authored publications is overwhelmingly unequal. This is in direct contrast to extant assumptions in research evaluation practice and professional science studies which presuppose equal contributions and do not adjust or weight publication and citation counts differentially by contribution. Such flawed methodology should be discontinued, as it unfairly disadvantages major contributors.

Introduction

A major open question in the science of science is how much, on average, do co-authors contribute to multi-author publications? And how are the number of co-authors of a paper and the position of an author's name in the author list related to the size of their contribution? The answers have far-reaching consequences for bibliometric research and the practice of research assessments of individuals, working groups, departments, organizations, and countries. A validated method for the allocation of relative credit for joint work to the involved contributors that reflects as closely as possible their relative contributions is indispensable for fair assessments and valid basic research. But we do not presently know enough about the typical contribution patterns in scientific teamwork. Are co-authors' relative inputs mostly equal or so close to equal as to be indistinguishable from equality? Or are they unequal, and if so, how much?

Many different co-author credit attribution schemes (or counting methods) have been proposed (Gauffriau, 2021) but it is not well known which of them are used in research and practice. The only study to investigate counting method use in scientometric research is that of Larsen (2008). Larsen analyzed the 85 accepted contributions to two conferences of the International Society for Scientometrics and Informetrics which used some method of publication counting. His summary of the findings is:

> *It is obvious that in more than half of the cases the information given on counting methods is insufficient. Whole counting is probably the dominating method but in more than half of the cases there is insufficient information to establish that whole counting was used. There is a nearly complete lack of arguments for the use of this method. (Larsen, 2008, p. 238)*

Only 31 % of the papers reported the applied method and a mere 6 % gave any justification for their choice. Several papers had problems with non-additive results. Non-additive results occur because, with whole counting, each contributing author is allocated one whole publication unit such that a paper with two authors is counted as two publications in total. Consequently, counts of all authors' publications are always greater than the true number of papers. Besides whole counting, the only other frequently used method at the time of Larsen's study was equal fractional counting, which is somewhat of an unofficial standard method in professional bibliometrics (Waltman, 2016). Equal fractional counting means that each author of a paper by n co-authors receives an equal share of credit of $1/n$. Equal fractional counting is preferred by professionals over whole counting (also called full counting) as it does not lead to inflated counts due to non-additivity when sums of participating units, such as authors and countries, are calculated to obtain total values.
Some researchers have noted a lack of empirical data to substantiate a decision to use a particular counting or credit attribution method (e.g., Petersen, Wang, & Stanley, 2010, p. 3). Korytkowski & Kulczycki (2019), after comparing several counting methods, conclude:

> *We have shown how different variants of publication counting methods influence the rankings. We could construct other variants, but it will not make our task, i.e. selecting the proper way of counting, any easier, because there is no external and objective reference point. (p. 815)*

However, there is at least some informative evidence on co-author contribution patterns. Evidence from qualitative studies in the sciences has accumulated, indicating that contribution-based name ordering is common (Knorr Cetina, 1999, p. 167; Laudel, 2001, p. 776; 2002, p. 11; Müller, 2012, pp. 301–303).
The most directly relevant and valid evidence comes from empirical studies of quantitative contribution estimation of authors themselves. Research on authors' own claims and statements of their relative contributions to co-authored work showed that contribution-based author name ordering is common and contributions are mostly unequal (Ali, 2021; Donner, 2020; Slone, 1996). But work in this approach has used quite small and unrepresentative samples. These scattered results are corroborated by a survey of active researchers from the UK, which found that:

> *The listing of authors in order of contribution (with first author providing the greatest contribution) is the most frequent practice in most disciplines except for the humanities where alphabetical order is the norm. But it is notable that in physical sciences, mathematics and social sciences alphabetical ordering*

> *and ordering by contribution are almost equally common. (Research Information Network, 2009, p. 26).*

Possible answer choices were, however, not mutually exclusive and several practices per discipline were commonly indicated.

Various co-author credit allocation methods, or bibliometric counting methods, have been proposed and the choice of method is important because the methods lead to very different results (Abramo, D'Angelo, & Rosati, 2013b, 2013a; Chudlarský, Dvořák, & Souček, 2014; Egghe, Rousseau, & Van Hooydonk, 2000; Gauffriau & Larsen, 2005; Korytkowski & Kulczycki, 2019; Moed, 2000). At higher levels of aggregation, such as countries, the differences between methods manifest primarily in citation impact indicator values rather than publication sums (Huang, Lin, & Chen, 2011; Lin, Huang, & Chen, 2013).

Despite their validity deficits, full counting and equal fractional counting still remain bibliometric standard methods, as no consensus has emerged on which of various more sophisticated credit allocation methods is most appropriate (Gauffriau, 2021; Ioannidis et al., 2007; Põder, 2022). In fact, much of the professional literature has confined itself to comparisons between only full counting and equal fractional counting (Gauffriau & Larsen, 2005; Korytkowski & Kulczycki, 2019; Põder, 2022; Stock, Dorsch, Reichmann, & Schlögl, 2023; Thelwall et al., 2023), which both assume and imply equal contribution of all co-authors. It is thus crucial to investigate the measurement validity of counting methods with respect to co-author contributions, that is, to study which of the methods is in closest agreement with actual co-author contributions, insofar as these can be quantified and collected.

Here we address the persistent problem of a lack of knowledge on empirical patterns of contributions to joint publications by co-authors (Korytkowski & Kulczycki, 2019; Moed, 2000; Narin et al., 1976; Price, 1981) by an analysis of patterns of quantified contribution with respect to papers' co-author counts and the position of co-authors' names in the author list in a large-scale data set of author-provided percentage contribution claims.

Methods and data

In this study we analyze a large dataset of author contribution statements for co-authored scientific and scholarly publications. These data were collected in the application process for two funding programs of the Tri-Council, three Canadian government research funding agencies. These are the Vanier Canada Graduate Scholarships, for prospective doctoral students, and the Banting Postdoctoral Fellowships. Both programs offer attractive conditions of fully financed three year (Vanier) and two year (Banting) research positions and are correspondingly highly selective. The two programs are administered by three funding agencies, each responsible for one broad area of research: CIHR/IRSC is responsible for health research, NSERC/CRSNG for the natural sciences and engineering, and SSHRC/CRSH covers the social sciences and humanities. For both programs, applicants submit a comprehensive application dossier which is the basis for the decisions of selection committees at the three funding organizations, which rank

applications according to the criteria of academic excellence, research potential, and leadership in the case of Vanier and the criteria of research excellence and leadership, quality of proposed research program, and institutional commitment and demonstrated synergy in the case of Banting. Each agency awards a similar number of scholarships and fellowships for a total of 166 Vanier and 70 Banting recipients annually.

When applying to either program, applicants are required to submit a publication list and to state their own contribution to all publications. This is done by filling a "Contribution Percentage" field that provides a dropdown menu from which applicants have to select a contribution range starting from "0-10" in increments of 10 %. Next to that field is a help tab that can be toggled when clicked which provides the following text: *"Based on your contribution role, indicate the approximate percentage (%) of work you contributed towards this publication, as a proportion of the total work contributed to this publication by all authors/contributors"*.

These publication contribution claims are the primary data for this study. Additionally we use metadata of the applicants' publications and socio-demographic and process variables from the application and administration system. Applicants submitted their contribution claims privately and under confidentiality. They only estimated their own, not all co-authors', contributions. The other co-authors' assessments of their own or the applicants' contributions were not collected. Thus, applicants made submissions with a presumably very low expectation that any co-author would see their claims. While co-authors might occasionally be reviewer panel members, in such cases they would not rate an applicant because of conflict of interest regulations. Because of these specific conditions and because of the well-established cognitive bias of overestimation of one's own contributions to teamwork relative to that of others (Broad, 1981; Caruso, Epley, & Bazerman, 2006; Herz, Dan, Censor, & Bar-Haim, 2020; Ilakovac, Fister, Marusic, & Marusic, 2007) we anticipate that applicants on the whole overstate their contributions. We make the assumption that this overestimation is independent of the number of co-authors of a paper and the applicant's author position on a paper, such that in effect their claimed overestimated contribution is proportional to their unobserved true contribution.

Contribution claims in Vanier and Banting applications have to be submitted by choosing a value range for one's own contribution from the ten ordered category ranges '1-10 %', '11-20 %', …, '91-100 %'. Researchers who apply are not told how and if their declared percent contributions will be used to evaluate their applications. We use these ordinal categorical data directly in rank-correlational analysis but transform them to their mid-range values (5, 15, ..., 95) for other analyses which require numerical data. Authors are often able to make a quantitative estimate of their own and co-authors' relative contributions to a common publication (Ali, 2021; Donner, 2020), although the uncertainty of such estimates is presumably substantial – how large remains an open issue for further research. As the data entry categories in this case are relatively fine-grained, inaccuracies of the mid-range point estimates with respect to the unknown true values necessarily have to be small.

The publication co-author contribution data include all applications for the years 2016 to 2021 independent of funding success: those which were not funded, those offered funding, and withdrawn ones. This time range was selected because this is the competition years for which the percent contributions were available, at the time of the study. This is not a random representative sample of the global community of researchers. It is restricted to early career researchers with ambitions to start or continue a research career.

This data set consists of 46,910 percentage contribution claims in 6,219 applications submitted by 5,547 unique applicants to one of the funding programs. Table 1 shows the sample sizes by agency. Additional information on the relevant publications was retrieved from CrossRef. For this, the CrossRef API was queried with the free text reference (including authors, date of publication, title, venue, volume, issue, publisher and page range). We retrieved the top 5 candidates, and used a python script (adapted from this repository: https://github.com/CrossRef/reference-matching-evaluation) to provide custom weights, and picked the candidate publication record with the highest score to get the DOI of each reference. In order to get the exact position of each applicant's name in the author list, we used a fuzzy matching approach based on the author list entered for each publication by the applicant.

Table 1. Sample overview.

Agency	Program	Number of applications	Number of unique applicants	Number of contribution claims
CIHR	Banting	1,183	1,048	12,841
NSERC	Banting	1,248	1,154	13,621
SSHRC	Banting	1,032	925	6,991
CIHR	Vanier	1,028	929	5,807
NSERC	Vanier	890	849	3,336
SSHRC	Vanier	800	745	2,511

We compare the empirically observed values of claimed percent contribution to the predictions that a selection of bibliometric counting methods make. We chose methods which divide one unit of publication credit such that the parts sum to 1.0, modified here to match the empirical data by multiplying by 100 to get percent values. We only chose methods which do not depend on choosing free parameters. We included equal fractional counting as the current standard method of professional bibliometrics and competing methods which divide the publication unit unequally according to different principles. Some propose monotonically decreasing credits as the author position increases, others propose different higher values for the last or later authors in the byline. In general, all the alternative methods were proposed with the intention to better reflect actual relative author contributions while avoiding inflating publication counts by yielding credit shares for one publication which sum

to more than unity. Despite this intention, they were not validated with empirical criterion data so far. The following methods are compared:

- Equal fractional counting. First suggested by Price & Beaver (1966).
- Harmonic counting. Proposed originally by Hodge & Greenberg (1981), re-introduced and empirically studied by Hagen (2008).
- Harmonic parabolic counting. Proposed in Aziz & Rozing (2013).
- Arithmetic counting. Proposed by Kalyane & Vidyasagar Rao (1995) and van Hooydonk (1997).
- 'Proportional' method of Howard, Cole & Maxwell (1987).
- Geometric count. Proposed by Egghe, Rousseau & van Hooydonk (2000).
- DFG (2004) 'rule of thirds'. Proposed to weight JIF points in performance based funding systems of German medical faculties and still used frequently for this purpose (Aman & van den Besselaar, 2024). The two-author case was not specified in the document but we split the credit evenly between both authors.

Further information and calculation formulas can be found in the respective cited references.

Results

Descriptive analysis

Figure 1 presents the average values of the claimed percent contributions for publications with two to five and ten co-authors (converted mid-range numerical values) and the values that the studied bibliometric counting methods give for the same input combinations. The error bars in panel *a* indicate 95 % confidence intervals for the means. Several notable observations can be made from the empirical contributions claims in panel *a*.

First, the average size of the claimed contribution in multi-author papers depends strongly on the author position. For instance the first author in a two-author papers on average claims 79 % while the second author claims 49 %. Second, these average claims do not add up to 100 %, thus, overestimation is confirmed. Third, size of claims only weakly depends on author count. The claims for first author, for instance, are all close to each other, although the claims decrease slightly with increasing author count. Fourth, the decrease in claimed contributions with increasing author position is not linear and flattens out while for papers of four and more authors we can discern a clear last-author effect such that this position's claims are higher than that of the preceding position. Fifth, confidence intervals for the means are small, indicating that there is close agreement on the typical claims across applicants contingent on author count and position. Comparing this pattern with those for the seven chosen bibliometric counting methods for the same author count and position data in panels *b* to *h*, none of the methods seems to be a very good approximation – with the pattern of equal fractional counting being obviously inconsistent with the empirical results. This comparison is continued in the following correlational analysis.

Correlation analysis

Table 2 shows the results of the correlational analysis of selected bibliometric counting methods for the empirical data. This excludes the data for single-author publications, as all methods credit them with a value of 100 %. The table contains Pearson correlation coefficients for data transformed to numerical figures using the range midpoint values, Spearman rank correlation coefficients for the untransformed original data (both with 95 % confidence intervals), and average absolute deviations between counting method values and transformed empirical data. Note that the empirical data is affected by overestimation which puts some unknown upper limit on possible correlations and a lower limit on average absolute deviation. We find it makes little difference whether we use original data and rank correlation or numeric estimates and Pearson correlation. For the whole dataset, geometric counting, harmonic counting, and the method of Howard, Cole, & Maxwell (1987) show the highest correlations with rank correlations of $\rho \approx 0.75$ each. Among these three, geometric counting has the smallest average absolute error with a misestimation of 20 percentage points (pp). Arithmetic counting, harmonic parabolic counting, and the method of DFG (2004) show rank correlations between 0.63 and 0.66. Equal fractional counting is aligned worst with the empirical data: $\rho = 0.40$, average absolute deviation: 32.5 pp. These results are consistent across major domains of research as the results disaggregated by agency show. Notably, equal fractional counting also does poorly in the social sciences and humanities, a domain with lower co-author numbers. Switching from equal fractional counting to, say, geometric counting, bibliometricians and research evaluators can reduce the average error in co-authored publication contribution estimation from 32.5 to 20.2 pp, which is a 38 percent relative improvement.

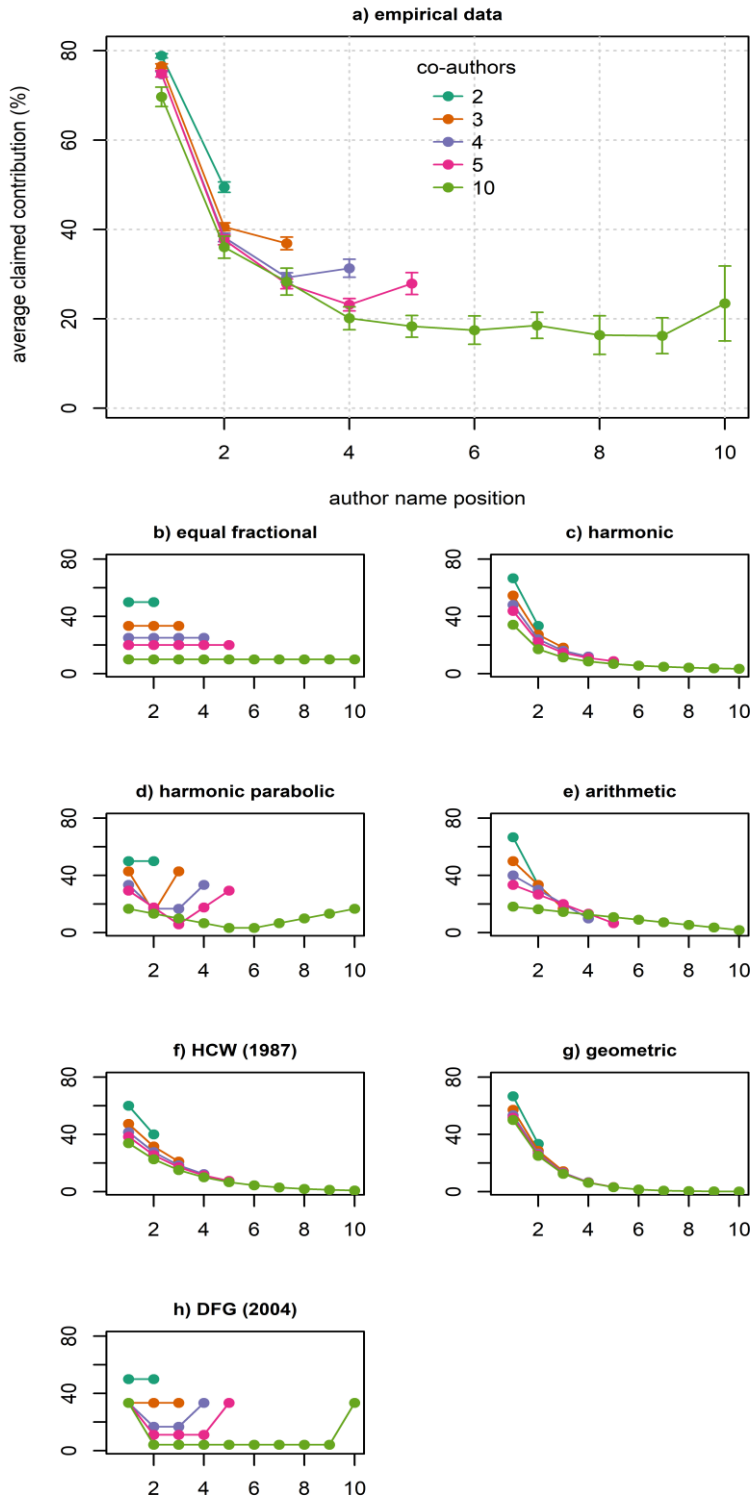


Figure 1. Average percent contribution to co-authored publications by author count and author position for empirical data and various bibliometric counting methods.

Table 2. Correlations of bibliometric counting methods with empirical contribution claims data.

Data set	bibliometric counting method	Pearson r	Spearman ρ	avg. abs. deviation
all science domains, N=37,157	equal fractional	0.40 (0.39, 0.40)	0.40 (0.39, 0.41)	32.5
	harmonic	0.76 (0.75, 0.76)	0.75 (0.75, 0.75)	22.7
	arithmetic	0.61 (0.60, 0.62)	0.64 (0.63, 0.65)	25.8
	Howard, Cole, Maxwell (1987)	0.74 (0.73, 0.74)	0.74 (0.74, 0.74)	24.3
	geometric	0.79 (0.79, 0.80)	0.75 (0.75, 0.76)	20.2
	DFG (2004)	0.68 (0.67, 0.68)	0.66 (0.66, 0.67)	29.6
	harmonic parabolic	0.61 (0.60, 0.62)	0.63 (0.62, 0.64)	30.6
CIHR - health research, N=16,379	equal fractional	0.39 (0.37, 0.40)	0.39 (0.38, 0.41)	33.1
	harmonic	0.77 (0.76, 0.77)	0.77 (0.76, 0.78)	23.1
	arithmetic	0.59 (0.58, 0.60)	0.63 (0.62, 0.64)	26.8
	Howard, Cole, Maxwell (1987)	0.75 (0.75, 0.76)	0.76 (0.76, 0.77)	24.3
	geometric	0.81 (0.80, 0.81)	0.77 (0.77, 0.78)	20.1
	DFG (2004)	0.73 (0.72, 0.73)	0.72 (0.71, 0.72)	29.4
	harmonic parabolic	0.63 (0.62, 0.64)	0.67 (0.66, 0.68)	31.0
NSERC - natural sciences and engineering research, N=15,440	equal fractional	0.44 (0.42, 0.45)	0.43 (0.42, 0.44)	33.8
	harmonic	0.78 (0.77, 0.78)	0.75 (0.74, 0.76)	23.3
	arithmetic	0.64 (0.63, 0.65)	0.66 (0.65, 0.67)	26.3
	Howard, Cole, Maxwell (1987)	0.76 (0.75, 0.77)	0.74 (0.74, 0.75)	25.3
	geometric	0.81 (0.80, 0.81)	0.75 (0.74, 0.76)	20.9
	DFG (2004)	0.70 (0.69, 0.71)	0.68 (0.67, 0.68)	31.2
	harmonic parabolic	0.65 (0.64, 0.66)	0.65 (0.64, 0.66)	31.7
SSHRC - social sciences and humanities research, N=5,338	equal fractional	0.32 (0.30, 0.35)	0.33 (0.30, 0.35)	26.6
	harmonic	0.68 (0.67, 0.70)	0.68 (0.66, 0.69)	20.0
	arithmetic	0.58 (0.57, 0.60)	0.61 (0.60, 0.63)	21.4
	Howard, Cole, Maxwell (1987)	0.66 (0.64, 0.67)	0.66 (0.64, 0.67)	21.0
	geometric	0.72 (0.71, 0.73)	0.68 (0.67, 0.70)	19.0
	DFG (2004)	0.49 (0.47, 0.51)	0.47 (0.44, 0.49)	25.7
	harmonic parabolic	0.48 (0.46, 0.50)	0.46 (0.44, 0.48)	26.1

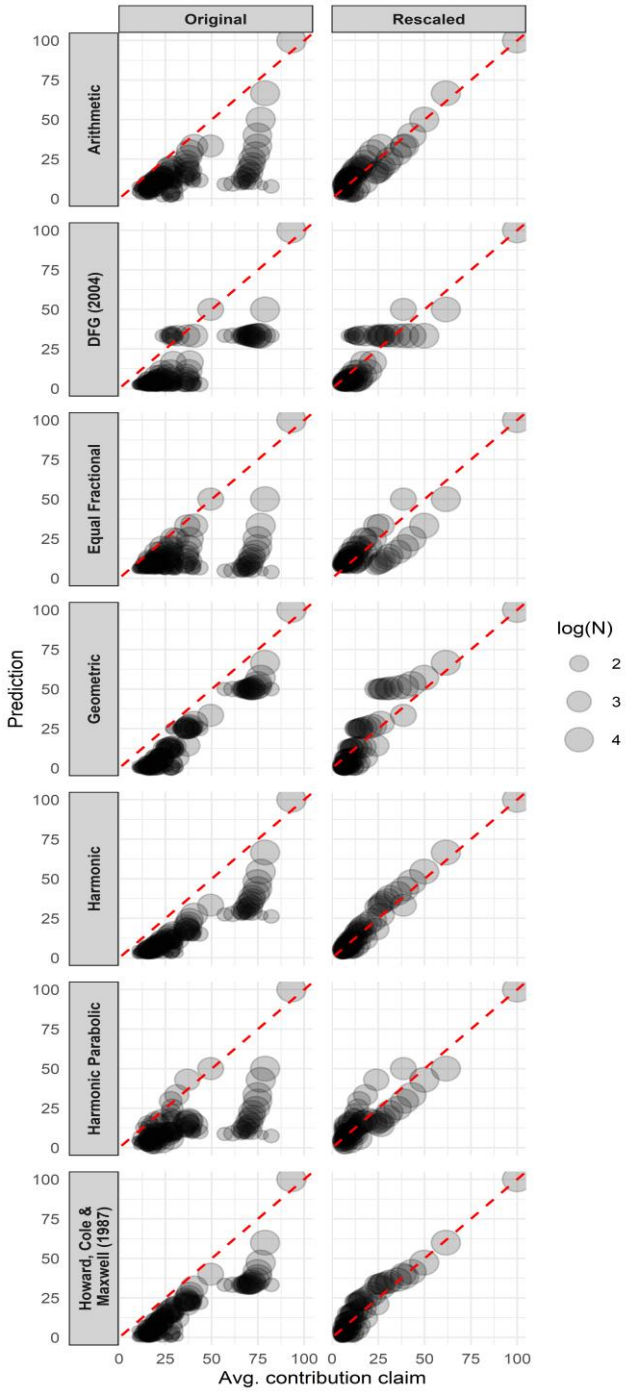


Figure 2. Comparison of empirical contribution data with bibiometric method predictions.

Note: Average values across combinations of author number and position displayed and scaled by log(N). Left, mid-range numerical values. Right, mid-range numerical values rescaled to sum to 100 % for each author count.

Figure 2 visualizes the comparisons of empirical data and counting method values for the calculated average values of contribution claims for each combination of author count and author position. On the left are the full data as transformed to numerical values, which include the overestimations. On the right we have removed the overestimations by rescaling such that the total sum for each author count data subset equals 100 %. For example, the contribution claim averages for first to third authors of three-author papers are 77, 41, 37 %. These were rescaled by the same factor to values of 50, 26, 24 %. This is only possible for subsets of the data for which enough observations for each author position are available, thus the right side plots only show the data for up to 12-author papers while the left side plots show more data. The figure indicates that some counting methods exhibit biased estimates in specific ranges. For example, the DFG (2004) method gives values which sum to 33 % to all first and last authors, these are mostly higher or lower according to the empirical data. The predicted values of equal fractional counting and harmonic parabolic counting are either too low or too high across most of the range. The method of Howard et al. (1987) and harmonic counting show very little bias.

Discussion

We have studied a large-scale dataset of percentage contribution claims by authors of co-authored scientific papers. The primary pattern that characterizes this data is profound inequality of contributions within one paper. As a first approximation, the author order tracks contribution order from most to least. An initial steep descent from first to middle authors is followed by a tapering off into a flat stretch, and, depending on author count, a final upturn for the last-author position. This empirical pattern of contributions resembles a ski jumping ramp, rather than the level plains which the equal contribution assumption of fractional counting implies.

Our findings indicate a misalignment between prevailing bibliometric methodology and real contribution patterns. Appropriate credit allocation is just as important for bibliometric research and research evaluation of higher aggregate units such as working groups, departments, and organizations as it is for individual co-authors. This is because the lower level units are mostly naturally nested within the higher level ones, such that credits for authors directly cascade up and can be aggregated to their affiliations by summation. The notable exception are multiple affiliations of a single author, which requires special handling. This not only goes for publication credit but is also relevant for citation analysis where co-author contribution shares are natural weights for fairly apportioning citation impact to co-authors and their affiliations.

In order to more closely reflect the actual contributions of co-authors, users of bibliometrics should phase out full and equal fractional counting and use counting methods that have been shown to agree much more closely with empirical contribution data in this study as these have higher validity. These may be the harmonic counting (Hagen, 2008; Hodge & Greenberg, 1981), geometric counting (Egghe et al., 2000), or the “proportional” method of Howard et al. (1987) or newly devised methods which align with actual contributions even better.

Acknowledgments

The views expressed herein are solely those of the authors and do not necessarily reflect those of CIHR.

References

Abramo, G., D'Angelo, C. A., & Rosati, F. (2013a). Measuring institutional research productivity for the life sciences: The importance of accounting for the order of authors in the byline. Scientometrics, 97, 779–795. https://doi.org/10.1007/s11192-013-1013-9

Abramo, G., D'Angelo, C. A., & Rosati, F. (2013b). The importance of accounting for the number of co-authors and their order when assessing research performance at the individual level in the life sciences. Journal of Informetrics, 7(1), 198–208.

Ali, J. R. (2021). Quantitative author inputs to earth science research publications: Survey results, insights and potential applications. Geological Magazine, 158(6), 951–963. https://doi.org/10.1017/S0016756820000916

Aman, V., & Van den Besselaar, P. (2024). Authorship regulations in performance-based funding systems and publication behaviour–a case study of German medical faculties. Journal of Informetrics, 18(2), 101500. https://doi.org/10.1016/j.joi.2024.101500

Aziz, N. A., & Rozing, M. P. (2013). Profit (p)-index: The degree to which authors profit from co-authors. PLoS One, 8(4), e59814. https://doi.org/10.1371/journal.pone.0059814

Broad, W. J. (1981). The publishing game: Getting more for less: Meet the least publishable unit, one way of squeezing more papers out of a research project. Science, 211(4487), 1137–1139. https://doi.org/10.1126/science.700819

Caruso, E. M., Epley, N., & Bazerman, M. H. (2006). The costs and benefits of undoing egocentric responsibility assessments in groups. Journal of Personality and Social Psychology, 91(5), 857. https://doi.org/10.1037/0022-3514.91.5.857

Chudlarský, T., Dvořák, J., & Souček, M. (2014). A comparison of research output counting methods using a national CRIS–effects at the institutional level. Procedia Computer Science, 33, 147–152. https://doi.org/10.1016/j.procs.2014.06.024

DFG. (2004). Empfehlungen zu einer Leistungsorientierten Mittelvergabe (LOM) an den Medizinischen Fakultäten. Stellungnahme der Senatskommission für Klinische Forschung [Position paper]. Retrieved from https://www.dfg.de/resource/blob/169106/eb4c72d6c6514800b4e2c83cf6e7641b/stellungnahme-klinische-forschung-04-data.pdf

Donner, P. (2020). A validation of coauthorship credit models with empirical data from the contributions of PhD candidates. Quantitative Science Studies, 1(2), 551–564. https://doi.org/10.1162/qss_a_00048

Egghe, L., Rousseau, R., & Van Hooydonk, G. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. Journal of the American Society for Information Science, 51 (2), 145–157. https://doi.org/10.1002/(SICI)1097-4571(2000)51:2%3C145::AID-ASI6%3E3.0.CO;2-9

Gauffriau, M. (2021). Counting methods introduced into the bibliometric research literature 1970–2018: A review. Quantitative Science Studies, 2(3), 932–975. https://doi.org/10.1162/qss_a_00141

Gauffriau, M., & Larsen, P. O. (2005). Counting methods are decisive for rankings based on publication and citation studies. Scientometrics, 64, 85–93. https://doi.org/10.1007/s11192-005-0239-6

Hagen, N. T. (2008). Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis. PLoS One, 3(12), e4021. https://doi.org/10.1371/journal.pone.0004021

Herz, N., Dan, O., Censor, N., & Bar-Haim, Y. (2020). Authors overestimate their contribution to scientific work, demonstrating a strong bias. Proceedings of the National Academy of Sciences, 117(12), 6282–6285.

Hodge, S. E., & Greenberg, D. A. (1981). Publication credit. Science, 213(4511), 950–950. https://doi.org/10.1126/science.213.4511.950.b

Howard, G. S., Cole, D. A., & Maxwell, S. E. (1987). Research productivity in psychology based on publication in the journals of the american psychological association. American Psychologist, 42(11), 975.

Huang, M.-H., Lin, C.-S., & Chen, D.-Z. (2011). Counting methods, country rank changes, and counting inflation in the assessment of national research productivity and impact. Journal of the American Society for Information Science and Technology, 62(12), 2427–2436. https://doi.org/10.1002/asi.21625

Ilakovac, V., Fister, K., Marusic, M., & Marusic, A. (2007). Reliability of disclosure forms of authors' contributions. CMAJ, 176 (1), 41–46. https://doi.org/10.1503/cmaj.060687

Ioannidis, J. P., Patsopoulos, N. A., Kavvoura, F. K., Tatsioni, A., Evangelou, E., Kouri, I., . . . Liberopoulos, G. (2007). International ranking systems for universities and institutions: A critical appraisal. BMC Medicine, 5, 1–9. https://doi.org/10.1186/1741-7015-5-30

Kalyane, V., & Vidyasagar Rao, K. (1995). Quantification of credit for authorship. ILA Bulletin, 30(3-4), 94–96.

Knorr Cetina, K. (1999). Epistemic cultures: How the sciences make knowledge. Harvard University Press.

Korytkowski, P., & Kulczycki, E. (2019). Publication counting methods for a national research evaluation exercise. Journal of Informetrics, 13(3), 804–816. https://doi.org/10.1016/j.joi.2019.07.001

Larsen, P. (2008). The state of the art in publication counting. Scientometrics, 77(2), 235–251. https://doi.org/10.1007/s11192-007-1991-6

Laudel, G. (2001). Collaboration, creativity and rewards: Why and how scientists collaborate. International Journal of Technology Management, 22(7-8), 762–781. https://doi.org/10.1504/IJTM.2001.002990

Lin, C.-S., Huang, M.-H., & Chen, D.-Z. (2013). The influences of counting methods on university rankings based on paper count and citation count. Journal of Informetrics, 7(3), 611–621. https://doi.org/10.1016/j.joi.2013.03.007

Moed, H. F. (2000). Bibliometric indicators reflect publication and management strategies. Scientometrics, 47(2), 323–346. https://doi.org/10.1023/A:1005695111622

Müller, R. (2012). Collaborating in life science research groups: The question of authorship. Higher Education Policy, 25(3), 289–311. https://doi.org/10.1057/hep.2012.11

Narin, F. (1976). Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity. Computer Horizons Cherry Hill, NJ.

Petersen, A. M., Wang, F., & Stanley, H. E. (2010). Methods for measuring the citations and productivity of scientists across time and discipline. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics, 81(3), 036114. https://doi.org/10.1103/PhysRevE.81.036114

Põder, E. (2022). What is wrong with the current evaluative bibliometrics? Frontiers in Research Metrics and Analytics, 6, 824518. https://doi.org/10.3389/frma.2021.824518

Price, D. de S. (1981). Multiple authorship. Science, 212(4498), 986–986. https://doi.org/10.1126/science.212.4498.986-a

Price, D. J. de S., & Beaver, D. deB. (1966). Collaboration in an invisible college. American Psychologist, 21(11), 1011–1018. https://doi.org/10.1037/h0024051

Research Information Network. (2009). Communicating knowledge: How and why UK researchers publish and disseminate their findings [Report].

Slone, R. M. (1996). Coauthors' contributions to major papers published in the AJR: Frequency of undeserved coauthorship. American Journal of Roentgenology, 167(3), 571–579. https://doi.org/10.2214/ajr.167.3.8751654

Stock, W. G., Dorsch, I., Reichmann, G., & Schlögl, C. (2023). Labor productivity, labor impact, and co-authorship of research institutions: Publications and citations per full-time equivalents. Scientometrics, 128(1), 363–377. https://doi.org/10.1007/s11192-022-04582-5

Thelwall, M., Kousha, K., Makita, M., Abdoli, M., Stuart, E., Wilson, P., & Levitt, J. (2023). Is big team research fair in national research assessments? The case of the UK research excellence framework 2021. Journal of Data and Information Science, 8(1), 9–20. https://doi.org/10.2478/jdis-2023-0004

Van Hooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. Journal of the American Society for Information Science, 48(10), 944–945. https://doi.org/10.1002/(SICI)1097-4571(199710)48:10%3C944::AID-ASI8%3E3.0.CO;2-1

Waltman, L. (2016). A review of the literature on citation impact indicators. Journal of Informetrics, 10(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007

https://doi.org/10.51408/issi2025_089

An Empirical Study on the Distributional Characteristics of Policy Citation Behaviors in Climate Action Policies

Zheng Xinman[1], Liu Xiwen[2]

[1]*zhengxinman@mail.las.ac.cn*, [2] *liuxw@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

Abstract

This study empirically analyzes the distributional characteristics of policy citation behaviors in climate action policies. By examining policy documents from different institutional sources, including intergovernmental organizations (IGOs), governments, and think tanks, the study finds that citation annotations are prevalent across all three types of institutions, with an overall usage rate of 87%. IGOs exhibit the highest utilization of citation annotations at 97.1%, followed by think tanks at 92.7%, and governments at 80.4%. The chi-square test confirms a statistically significant difference in citation annotation usage among these institutions. The study identifies six common types of citation annotations: footnotes, endnotes, bibliographies, in-text citations, captions, and hyperlinks. Footnotes and bibliographies are the most frequently used types across all policy sources, accounting for over 60% of total citations. However, preferences vary among institutions; IGOs favor captions, think tanks prefer bibliographies and in-text citations, while governments predominantly use footnotes and hyperlinks. Think tank policies exhibit the highest citation frequency, while government policies have a relatively lower rate. These findings shed light on the differences in citation behaviors among various policymaking institutions and provide insights into the Science-Policy Interface in climate action policies.

Introduction

Policy document citations are citations of external information within policy texts, similar to citation data in academic papers, and have a wide range of research value and applications. By analyzing policy document citations, it is possible to analyze the Science-Policy Interface (SPI), which refers to the interaction and mutual influence between scientific research and policy-making, or to assess the social impact of scientific research publications (Bornmann, 2016; Haunschild and Bornmann, 2017; Bornmann, 2022). In addition, quantitative analyses of policy document citations can enrich the research scope of public policy analysis by providing statistical data on how policy draws on external information (Newson, 2018), thereby expanding the research paradigm of public policy. The quantitative analysis of policy citation data, which reflects policy citation behavior, refers to how external information is used in policy texts. Therefore, understanding the specific ways of the distribution of policy citation behavior helps to better understand the connotation of policy citation data and promote the further development of policy citation research.

Current research on the distribution of policy citation behavior faces a triple challenge: limited data availability, unclear institutional variations and methodological constraints. First, the policy document citations used in the current

study mainly come from policy citation data provided by Altmetric and Overton databases, which suffer from data coverage bias (Bornmann, 2016; Bornmann, 2022; Tattersall, 2018; Maleki, 2022), and some policy document citation behavior may not be supported and identified by the databases due to unstructured features (Overton, 2019). The resulting distribution of policy document citations based on citation behavior data from these databases appears to be similar in nature to the distribution of academic citations (Szomszor and Adie, 2022), but conclusions such as the generally low proportion of citations from academic papers in policy texts lead to difficulties in determining whether policy citation behaviors are systematic practices or accidental manipulations, which, in turn, undermines the ability to ability to verify the universal law of policy citation behavior. Second, different policymaking institutions (e.g., governments and think tanks) show differences in the use of academic citations in policymaking; for example, during the COVID-19 pandemic, the tendency to cite science in policy documents seems to have been concentrated mainly within intergovernmental organizations (IGOs), such as the World Health Organization (WHO), and to a much lesser extent in national governments, which mainly consume science through intergovernmental organizations indirectly consume science (Yin, 2021). Such differences may stem from differences in institutional resource endowments or knowledge translation mechanisms, but existing studies have not yet been able to further reveal and quantify the extent of differences in cross-agency policy citation behavior, nor the selection mechanisms and influencing factors behind them, due to a lack of comparative analysis of cross-agency policy citation behavior. More critically, the unstructured characteristics of policy texts and the existence of multiple citation styles make the automated extraction of citation data challenging, and it is difficult to extract citation data directly from policy texts and consumes a large amount of labor costs (Newson, 2018), which leads to the reliance of existing studies on the distribution of policy citation behaviors on small samples of manual annotations (Newson, 2018; Yu, 2023), making the comparison of policy citation behavior differences across sources lacking data support. These three obstacles together constitute the “black box” of the policy citation behavior distribution problem - do policies in all fields follow the same citation pattern? What structural factors drive the heterogeneity of citation patterns?

Citation annotations in policy documents are an important basis for analyzing policy citation behavior, which can provide key clues for deciphering the “black box” problem in current research. Existing research shows that citation annotations in policy documents are mainly divided into two categories: one is the use of specific wording or quotation marks to cite external sources of information in the body or table headings, footnotes and endnotes, such as “based on”, “refer to” and other prompt words such as “see” (Huang, 2015; Overton, 2022), or quotation marks in the body text of policy documents (Ba, 2022); the other is referencing styles in common publications such as academic papers, such as footnotes, endnotes, hyperlinks, bibliographies, etc. (Newson, 2018; Yu, 2023). For example, Newson et al. found that approximately two-thirds of childhood obesity prevention policy documents issued by the New South Wales government in Australia between 2000 and 2015

contained references, and of these, more than one-third of the policies used footnotes, hyperlinks, or a combination of these forms (Newson, 2018).Yu et al. found that the standardized referencing style was the main form of reference when policies cite academic papers, as demonstrated by citing academic papers in the form of post-textual reference lists and including footnotes or endnotes in the body of the policy (Yu, 2023). It can be seen that policy citation annotations have the potential to analyze the distribution of policy citation behaviors.

The United Nations Sustainable Development Goals (SDGs) are complex global issues that involve a wide range of policymaking institutions and stakeholders, including governments (national and local), IGOs, the private sector, non-governmental organizations (NGOs) and academia. Among them, SDG 13 (Climate Action) has a wide range of impacts, covering a variety of fields such as environment, energy, economy, etc., and has produced a wealth of policy documents and scientific research results. As a common challenge faced by all mankind, the types of institutions that formulate climate action policies are many and numerous, and the degree of policy disclosure is high (Bornmann, 2022). As time progresses, global climate governance faces important challenges that call for more scientific policy development and greater citation of evidence, and therefore has the potential for generalized use of citation behaviors in its policies compared to policies on other topics, but current research based on policy citation databases suggests that climate action policies cite science at a low rate (Bornmann 2016) and it is not clear that policies from different sources have similar citation behavior. In addition, policymaking institutions have their own positions, and their processes and roles in policy development vary, which may lead to differences in whether and in what form references to sources of information are included in policy documents.

This study explores the prevalence and variability of the distribution of policy citation behaviors in the field by analyzing citation annotations in climate action policy documents. Three specific issues are analyzed: first, an analysis of policy availability, which explores the main policy sources and document styles of policies, as well as the availability of policies; second, an analysis of the prevalence of policy citation behaviors across different policy sources and formats; and finally, an examination of the differences in citation annotations choices and use across different policy sources. Based on these analyses, important support is provided for understanding the prevalence and differences in the distribution of policy citation behaviors.

Method

Policy Document Source Identification

We used the Overton policy document database to retrieve policy documents from different sources in the field of climate action for two reasons. The first is that the Overton database covers a wide range of institutional types of sources of policy documents. The other is that the Overton database maps policy documents to one or more of the SDGs. The Overton database defines a policy document very broadly as “documents written primarily by and for policy makers”. This idea is intended to

cover not only policy documents documenting the policy or legislation itself, but also documents intended to inform or influence decision-making (Szomszor and Adie, 2022). The policy scope of this paper is consistent with Overton's definition. This paper combines the SDG labels provided by the Overton database to select SDG 13 policy documents for climate action, totaling 22,352. These policy documents contain the types of Publication, Blog post, and Working paper. In this paper, we choose publication as the sample of policy citation annotations because it is a formally released document with a relatively standardized style, and we get 20,303 policy documents with three types of institutional sources, including 3,954 documents from IGOs, 10,330 documents from governments, and 6,019 documents from think tanks in various countries.

Policy Document Collection and Sampling

In order to understand the characteristics of policy citation annotations for each type of organization, this paper draws samples from the policies of each type of organization separately for fine-grained annotations by means of stratified sampling. Stratified sampling, also called type sampling, is a sampling method that divides the overall units into a number of types or strata according to their attribute characteristics, and then randomly selects sample units from the types or strata. Stratified sampling is characterized by the fact that the commonality between units in each type is increased through the delineation of types and strata, and it is easy to draw a representative survey sample. This method is suitable for the overall situation is complex, the difference between the units is large, more units, applicable to the application of this paper's scenario. The specific process of stratified sampling is to first calculate the sampling proportion of each institutional category, and for each category, multiply it by the total sample size to get the sample size that should be taken for that category. Random sampling is then performed in each category to ensure that the sample in each category is random. Rogers et al. consider the amount of literature data used for bibliometric analysis to be at least 200 . This study refers to this criterion and 200 policy documents were sampled to ensure that the sample size was sufficient for econometric analysis. The stratified sample yielded 38 IGO policies (proportion: 19%), 102 government policies (proportion: 51%), and 60 think tank policies (proportion: 30%).

Policy Citation Annotation Coding

The original text of the sampled policy documents was downloaded according to the URL provided in the Overton database. The coding yielded information about each policy document, including two categories of policy document basic information and policy citation annotations information. The basic information of the policy document includes the title of the policy document, the source country, the name of the source organization, the type of the source organization, the link to the original policy text, the date, the availability of the policy document, and the type of the policy document format. Most of the information comes from data items exported from the database. Policy document availability, policy document format type, and total number of pages in the policy document are manually coded, and are judged and

counted when the original policy text is downloaded. Policy citation annotation information includes information on whether it contains citation annotations, the type of citation annotations, and the number of times the citation annotations were used. Policy citation annotations were obtained manually by scanning the full text of each policy through a combination of manual identification and content analysis to find and record the types of citation annotations that appeared in the main text and appendices and the frequency of use of that type of citation annotation in the policy documents.

Determines whether the policy contains citation annotations

Considering the universality and consistency of citation formats in academic papers, this paper identifies and records citation annotation types based on the reference citation formats commonly used in academic papers. If a policy document contains at least one type of reference citation format, it is registered as "containing citation annotations." By summarizing the commonly used citation annotation types in various publications, along with their positions and forms, citation annotations can be classified into six types: Footnotes, Endnotes, Captions (below tables or figures), In-text citations, Bibliographies, and Hypertext Links. Among these, In-text citations differ from Footnotes, Endnotes, and Hypertext Links in the formatting of the markers used when directly quoting content within the text. In-text citations typically adopt a parenthetical format indicating the author-date next to the quoted content, formatted as (Author, Year), for example, (Smith, 2019). Different academic writing style guides (such as APA, MLA, Chicago, etc.) may exhibit some variations in the formatting of in-text citations. However, regardless of the citation style, it is advocated to provide basic information about the cited content within the text (such as the author's name, publication year, title of the article or book, etc.), enabling readers to accurately understand the source and context of the quoted content. Footnotes and Endnotes commonly use numerical or symbolic markers. A Hypertext Link, also known as a Hyperlink or simply a Link, is used in web pages or electronic documents to direct users to other pages, resources, or locations when clicked. Hypertext Links are usually presented in text form and are often highlighted by changing the color of the link text or by underlining it.

Determining the type of citation annotations

Check whether the policy documents contain six common types of annotations, such as “reference lists”, “in-text markup”, “footnotes”, “endnotes”, “notes below charts”, “hypertext links”, etc., and determine whether these types of annotations play the role of citation. “Six common types of annotations, including footnotes, endnotes, and hypertext links, were examined to determine whether or not they played a role in citation annotations. We take into account the cases where footnotes, endnotes, and notes underneath charts and tables may play a non-citation annotations role, such as terminology explanations only, and so on. Since the intent of this paper is to observe citation behaviors in policy documents, only annotation types that play a citation role are registered in this paper. When a certain annotation type provides external sources in the policy documents, it can be regarded as playing the role of citation, registering

this annotation type as the citation annotations of the policy, and taking the frequency of this annotation in the text as the frequency of the annotation use. When a certain type of annotations only explain the role, there is no citation annotations, do not register the type of annotations. If there is a citation of external sources of information in the text, but the citation annotations do not belong to the six common types of annotations, the type will be registered as other types.

Counting the frequency of use of citation annotations

In order to facilitate counting and reduce labor costs, the total number of times a certain type of citation annotations appear in a single text as the citation annotation frequency, without the need to distinguish one by one which content is a citation and which is an explanation, in order to ensure that the identification of which types of annotations play the role of citation, and greatly improve the efficiency of manual labeling.

Results

Policy documents accessibility

The integration of URLs obtained through sampling and the subsequent download of original policy documents yielded a comprehensive dataset comprising 55 think tank policies, 34 intergovernmental organization (IGO) policies, and 92 government policies, resulting in a total sample size of 181 policies. The sample encompasses contributions from 21 countries, 15 IGOs, and 39 think tanks, indicating a diverse and extensive range of sources contributing to policies related to Sustainable Development Goal (SDG) 13 on climate action.

Table 1. Distribution of Policy Sample Sources.

Source Type	*#countries/regions*	*#institutions*	*Top 3 institutions/countries by frequency*
IGO	——	15	UNEP, World Bank, FAO
Government	21	53	USA, EU, UK
Think Tank	13	39	USA, UK, Belgium, Germany

The vast majority of policy documents in the sample (90.5%) were publicly accessible through existing or archived websites, with 87% available in PDF format and 3.5% in HTML format. However, 19 documents were unavailable due to inaccessible web pages (e.g., “page not found” or “404 - file or directory not found”), lack of access rights, or misclassification as policy documents (e.g., conference proceedings unrelated to institutional policies). These 19 documents, which could not be retrieved or were deemed irrelevant, were categorized as “other,” accounting for 9.5% of the sample, as illustrated in the table below. The findings indicate that the availability of policies across different sources exceeds 80%, reflecting a relatively high level of accessibility. This availability rate is notably higher than that

of the Overton database, which itself surpasses the percentage of valid policy data in Altmetric.com (71%) (Yu H, 2023). These results underscore the robustness of the dataset and the comparative advantage of the Overton database in terms of policy data accessibility.

Table 2. Accessibility of policies from different sources.

Source Type	*PDF Format*	*Html Format*	*Unable to Obtain*
IGO	89.5%（n=34）	0.0%（n=0）	10.5%（n=4）
Government	89.2%（n=91）	1.0%（n=1）	9.8%（n=10）
Think Tank	81.7%（n=49）	10.0%（n=6）	8.3%（n=5）
Total	87.0%（n=174）	3.5%（n=7）	9.5%（n=19）

Distribution of climate action policy citation behaviors

The utilization rate of citation annotations in climate action policies: The analysis of citation annotation usage rates across policies from different institutional types revealed that, overall, 87% (n=181) of climate action policies included citation annotations. This high percentage underscores the prevalence of citation practices within climate action policies, suggesting that referencing and acknowledging sources is a common and integral aspect of policy development in this domain.

Usage rates of citation annotations for different institution types: In terms of the type of institution, the proportions of citation annotation policies originating from governments, think tanks, and intergovernmental organizations (IGOs) within the sample set of policies from their respective sources are 80.4%, 92.7%, and 97.1%, respectively. These figures indicate a higher prevalence of citation annotations in policies issued by these three types of institutions, as illustrated in the table below. Yin et al. posited that IGOs exhibit a more pronounced tendency to cite scientific research in their policy documents compared to national governments (Yin, 2021). The findings of this study corroborate this assertion, revealing that IGOs have the highest utilization of citation annotations, while governments have the lowest. This suggests that the policy citation behaviors, whether or not it pertains to scientific research, is more prevalent among IGOs than among governments.

Table 3. Citation annotation usage rate of policies from different sources.

Source Type	*With Citations*	*Without Citations*
IGO	80.4%	19.6%
Government	92.7%	7.3%
Think Tank	97.1%	2.9%
Total	87%	13%

After conducting the chi-square test, the obtained test values for Pearson's chi-square and the likelihood ratio were 0.016 and 0.010, respectively. Both of these values are less than the conventional significance level of 0.05 (p-value). This indicates that there is a statistically significant relationship between the type of institution and the rate of policy use of citation annotations. Specifically, there is a significant difference in the rate of policy citation annotations usage among governments, think tanks, and IGOs. As detailed in the table below, IGOs and think tanks exhibit a greater inclination to utilize citation annotations in their policy documents compared to governments.

Table 4. Chi-Square Tests for the Relationship Between "Source Type" and "Citation Annotation Usage Rate".

Test Type	*Value*	*Degrees of Freedom*	*Asymptotic Significance (2-sided)*
Pearson Chi-Square	8.289	2	0.016
Likelihood Ratio	9.196	2	0.010
Linear-by-Linear Association	7.740	1	0.005
Valid Cases	181	——	

Utilization of citation annotations across various file format types: The statistics presented in Figure 1 provide insights into the utilization of policy citation annotations across different document format types. Upon analyzing Figure 1, it becomes evident that the proportion of policy documents employing citation annotations is notably higher in both PDF and HTML formats. This suggests that policy citation behaviours are prevalent in documents where citations are explicitly manifested.

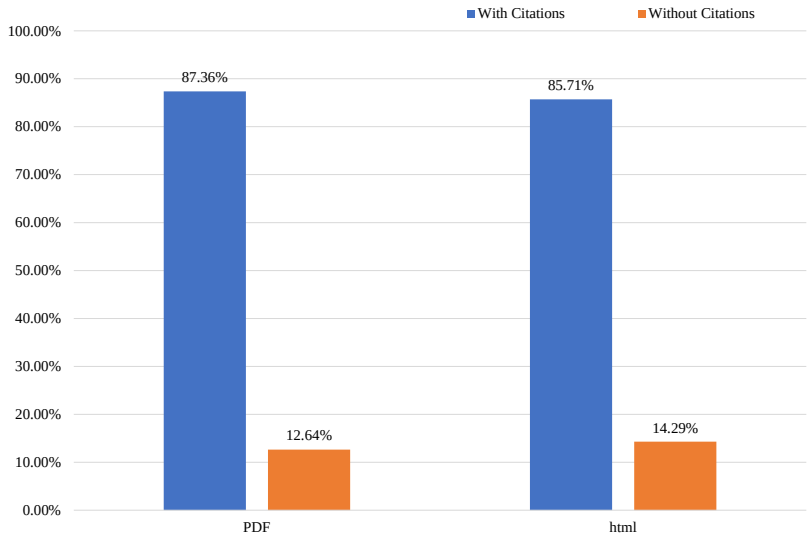


Figure 1. Proportion of Citation Annotations in Sample Policies.

Variations in citation behaviors across policy sources

Citation Frequency in Policy Documents

An analysis of the cumulative distribution of citation counts for policies originating from diverse sources unveils a pronounced imbalance in the frequency of policy citations among these sources. Specifically, it is observed that 27% of government policies account for 80% of the total citation counts, 30% of policies from IGOs contribute to 80% of the citations, and 37% of think tank policies are responsible for 80% of the citations. This distribution pattern indicates that think tank policies generally exhibit a higher citation frequency, whereas government policies demonstrate a relatively lower citation rate, highlighting disparities in policy-making practices across different types of organizations. This phenomenon, wherein a minority of policies garner the majority of citations, aligns with the Pareto Principle, which posits that in numerous instances, roughly 20% of the factors (policies) generate approximately 80% of the outcomes or impacts (citations).

Upon analysing the citation frequency (total count of citation annotations per policy) and citation density (number of citation annotations per page of policy) for individual policies across different source institution types, it is evident that think tank policies exhibit the highest average number of citation annotations. This is followed by policies from IGOs, with government policies trailing behind. Notably, there is a higher prevalence of outliers in the citation data for government policy documents, indicating a greater degree of variation in citation behaviours among government policies compared to those from think tanks and IGOs.

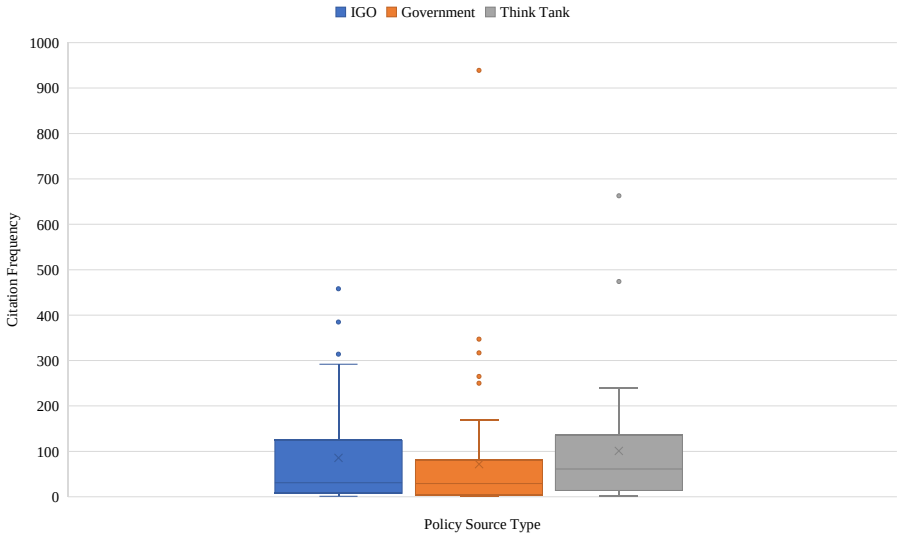


Figure 2. Distribution of Citation Frequency for Policies from Different Policy Source Type.

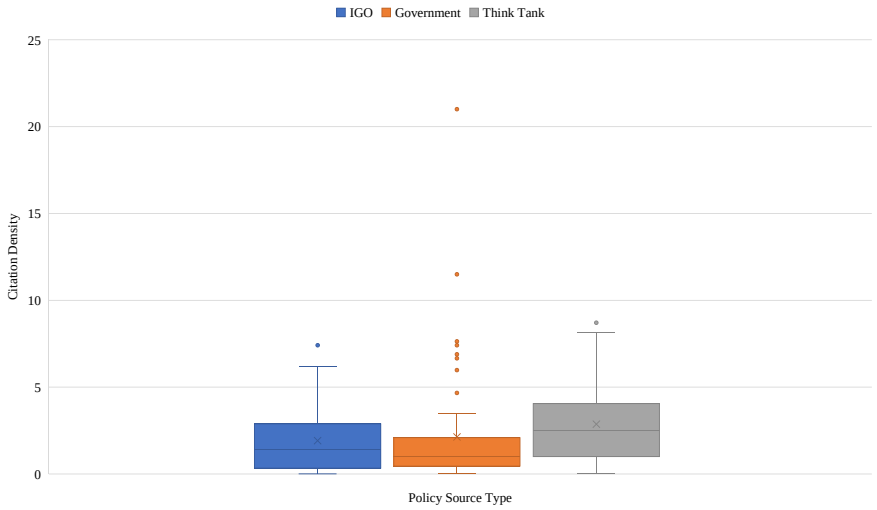


Figure 3. Distribution of Citation Density for Policies from Different Policy Source Type.

Usage of policy citation annotation types

Types of policy citation annotations: The annotation results reveal that there are a limited number of text box (Box) annotations that contain citation annotations within the policies. Upon examination, it becomes apparent that both the chart annotations and text box annotations share some similarities in their citation behaviour. Specifically, citations in these two cases typically occur in prominent and distinct locations within the text, and sometimes exhibit more independent citation patterns. In the text, citations are usually introduced using source or note prompt words. For the purpose of simplifying the analysis, this study groups these two types of

annotations into a single category, referred to as caption. By combining the common citation annotations found in policy documents with the types of annotations within the text, we ultimately identify six types of policy annotations: footnotes, endnotes, bibliographies, in-text citations, captions and hyperlinks.
In terms of the choice of citation annotation types, an analysis of the frequency and distribution of each annotation type within policy documents provides insights into the characteristics of policies with regard to their citation practices. The following table presents an overview of these findings:

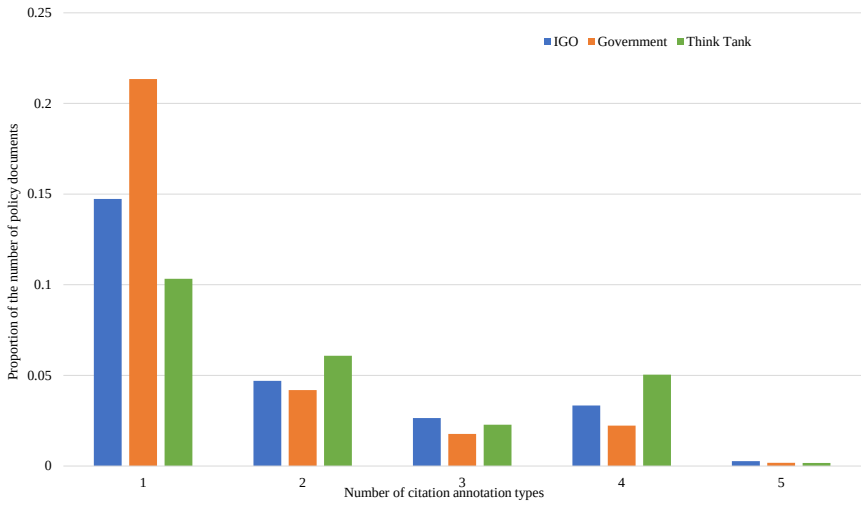


Figure 4. Number of citation annotation types used per policy (Normalized).

Analysing the provided figure, it is evident that the majority of policies utilize only 1 type of citation annotation. Additionally, there are policies that employ up to 5 types of citation annotations. Notably, policies containing fewer than 2 types of citation annotations constitute 60.5% of the policy document citations. This finding indicates that policies from different institutional types exhibit a common trend in their selection of citation annotations; specifically, they tend to use fewer than 2 types of citation annotations.
Analyse whether there is a common use of a certain citation annotations type across institutions. An examination of the types of citation markers contained in climate action policies was conducted, with the percentage of policies from each institution type utilizing each citation marker type calculated and presented in Table 2 below. As is shown in the table, over 40% of policies from all three types of institutions employed hyperlinks and footnotes, indicating that these are the most frequently used citation marker types in climate action policies. A chi-square test was employed for analysis, and the results revealed no significant differences ($P > 0.05$) in the usage rates of footnotes and hyperlinks among the three institution types. This result aligns with existing research on policy citation annotations by type of government agency. Newson (2018) noted that "more than one-third of policy documents do not list references in individual lists or appendices, but instead use footnotes, hyperlinks, or

a combination of these methods." This consistency in findings reinforces the notion that footnotes and hypertext links are widely accepted and utilized as effective means of citing sources in policy documents, particularly in the context of climate action policies.

Further compare whether there are significant differences in the types of policy citation annotations preferred by different institution types. Based on the observations, it appears that different types of institutions have distinct preferences when it comes to the types of citation annotations used in their climate action policies. Bibliographies and captions emerge as the most common citation annotation types for think tanks ($P<0.05$). IGOs favours using captions as citation annotations, which involve direct source citations at pictures, tables, and separate text boxes, ($P<0.05$). In contrast, government policies predominantly use footnotes and hypertext links as their most common citation annotations. Notably, governments are less likely to use endnotes compared to think tanks and IGOs ($p < 0.05$). Considering the varying primary modes of dissemination for policies among different institution types, there are also differences in the preferred types of citation markers. Different citation marker types serve distinct purposes and effects. For instance, the combination of bibliographies at the end of the document and in-text citations is the most common type in academic publishing, while hyperlinks are a convenient citation marker type for online publishing and dissemination.

Table 5. Comparison of the Proportion of Policies with Specific Citation Annotations by Source Type.

Citation Annotation Type	*Government*	*Think Tank*	*IGO*	*P Value*
Hyperlink	42.4	41.8	55.9	0.350
Footnotes	54.3	41.8	44.1	0.286
Endnotes	3.3	25.5	11.8	0.000
Captions	22.8	50.9	55.9	0.000
In-text citations	15.2	47.3	32.4	0.000
Bibliography	19.6	50.9	26.5	0.000

Common types of citation formats from different sources

Common citation annotation types in various policy sources. Considering that the citation content contained in in-text citations overlaps with bibliographies or endnotes, in-text citations were excluded from the analysis. The usage frequencies of the remaining citation annotation types were counted, resulting in a stacked percentage bar chart of citation annotation usage, as shown in the figure below. The results indicate that both footnotes and bibliographies account for 60% or more of the total citation annotation usages in policies from the three types of institutions.

From the perspective of usage frequency, footnotes and bibliographies are the most frequently used citation annotation types across policy texts from various sources.

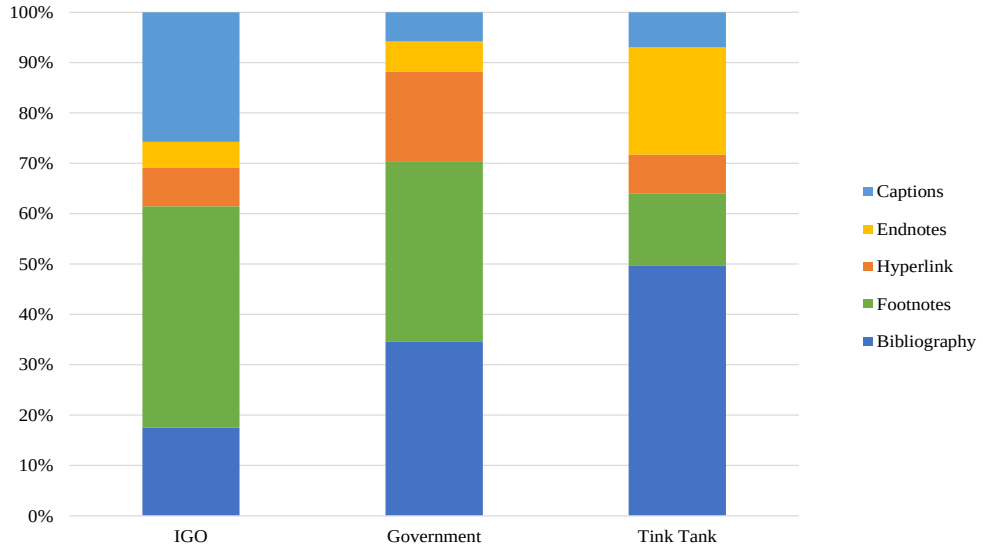


Figure 5. Percentage Stacked Bar Chart of Citation Annotation Usage.

Preferences in the usage frequencies of various citation annotation types. Upon analysing the figure below, it can be observed that the citation annotation types with the highest average usage frequencies in IGO policies are footnotes and bibliographies. In think tank policies, the citation annotation types with the highest average usage frequencies are in-text citations, bibliographies, and endnotes. Similarly, in government policies, the citation annotation types with the highest average usage frequencies are also in-text citations, bibliographies, and endnotes. Evidently, bibliographies emerges as the most frequently used citation annotation type across policies from all three sources.

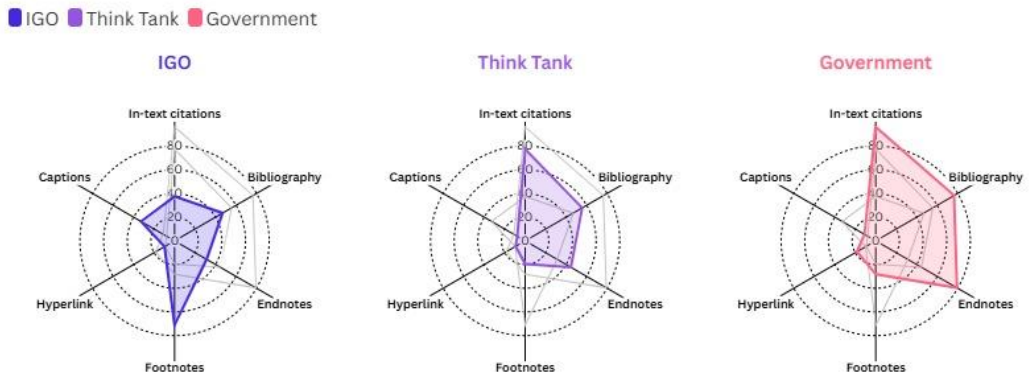


Figure 6. Average Usage Frequency of Each Annotation Type in Different Policy Sources.

Discussion

The sources of climate action policies are diverse, encompassing IGOs, governments, and think tanks across countries as the primary origins, and these policy documents are generally highly accessible, facilitating the direct extraction of policy citation data from the policy texts. Given the prevalence of policy citations in climate action policies, the direct extraction of such data holds significant importance for analyzing the sources of scientific evidence in the policy-making process, assessing policy impacts, and promoting the interaction between science and policy. Furthermore, with advancements in text mining and natural language processing technologies, it has become feasible to directly extract policy citation data from policy texts.

Policies from different sources exhibit significant variations in citation behavior. Think tanks and international governmental organizations (IGOs) tend to cite a wide range of literature to support their policy proposals, reflecting a strong commitment to evidence-based decision-making. In contrast, government agencies may cite relatively fewer references, as they often prioritize practical implementation and immediate effects. This disparity reflects differing approaches to evidence-based decision-making among various decision-making bodies: think tanks and IGOs emphasize the foundational role of scientific research in policy formulation, while government agencies focus more on the timeliness and operationalizability of policies. When promoting the interaction between science and policy, it is essential to fully consider the characteristics and needs of different decision-making bodies.

Limitation and Future Work

Limitation. This paper collects data from Overton, a process that inevitably involves a certain degree of selection bias regarding policy data sources, resulting in a higher likelihood of capturing only those policy data sources that are readily accessible. This limitation in data sources may imply that some non-public or hard-to-access policy documents are omitted, thereby affecting a comprehensive and in-depth analysis of policy citation behavior.

We employ explicit citation markings to explore policy citation behavior; however, the behavior itself is exceedingly complex. Extracting policy citation data solely from policies containing citation markings clearly cannot fully capture the entirety of policy citations. As policy documents often do not disclose whether specific evidence evaluation criteria have been applied, there exists the possibility that some studies, although utilized, are not explicitly cited. This constitutes a limitation of our research method.

Future Work. This study has conducted an in-depth exploration of policy accessibility and citation marking styles and preferences, laying a foundational groundwork that provides valuable experience and insights for the design of subsequent automated citation extraction methods. By understanding the avenues for obtaining policy documents and the diversity and preferences in citation markings, we can design algorithms and models more targetedly to enhance the accuracy and efficiency of automated citation extraction. This will significantly promote the large-scale acquisition and utilization of policy citation data, offering new tools and methods for policy research.

The extraction of policy citation data not only provides abundant material for research on the interaction between science and policy, but also enables the assessment of the impact of different types of publications on policy. By analyzing the sources, types, and frequencies of citations in policy documents, we can reveal which publications have exerted significant influence on policy formulation and how this influence occurs. This will facilitate a deeper understanding of the interaction between science and policy, providing scientific evidence for policymakers and policy-oriented references for publication editors and authors.

Comparing citation behaviors across policies in different domains is an important and intriguing issue. For instance, there may be significant differences in policy citations between clinical research and public health research. By conducting a comparative analysis of the citation characteristics of policies in these two domains, we can uncover the sources of scientific evidence and decision-making logic underlying policy formulation in different fields, as well as their demands and preferences for scientific research. This will contribute to a more comprehensive understanding of the nature and patterns of policy citation behavior, providing targeted suggestions and guidance for policymakers and researchers in various domains.

Acknowledgments

The policy document data were shared with us by Overton on September 11, 2023.

References

Ba, Z., Zhao, Y. C., Liu, X., & Li, G. (2022). Spatio-temporal dynamics and determinants of new energy policy diffusion in China: A policy citation approach. Journal of Cleaner Production, 376, 134270.

Bornmann, L., Haunschild, R., & Marx, W. (2016). Policy documents as sources for measuring societal impact: How often is climate change research mentioned in policy-related documents?. *Scientometrics*, *109*, 1477-1495.

Haunschild, R., & Bornmann, L. (2017). How many scientific papers are mentioned in policy-related documents? An empirical investigation using Web of Science and Altmetric data. *Scientometrics*, *110*, 1209-1216.

Huang, C., Su, J., Xie, X., Ye, X., Li, Z., Porter, A., & Li, J. (2015). A bibliometric study of China's science and technology policies: 1949–2010. Scientometrics, 102, 1521-1539.

Maleki, A., & Holmberg, K. (2022). Comparing coverage of policy citations to scientific publications in Overton and Altmetric. com: Case study of Finnish research organizations in Social Science. Informaatiotutkimus, 41(2–3), 92-96.

Newson, R., Rychetnik, L., King, L., Milat, A., & Bauman, A. (2018). Does citation matter? Research citation in policy documents as an indicator of research impact–an Australian obesity policy case-study. *Health Research Policy and Systems*, *16*, 1-12.

Overton. How does Overton find citation contexts? Overton. Retrieved December 7, 2024 from: https://help.overton.io/article/how-does-overton-find-citation-contexts/#supported-referencing-styles

Szomszor, M., & Adie, E. (2022). Overton: A bibliometric database of policy document citations. *Quantitative science studies*, *3*(3), 624-650.

Tattersall, A., & Carroll, C. (2018). What can Altmetric. com tell us about policy citations of research? An analysis of Altmetric. com data for research articles from the University of Sheffield. *Frontiers in research metrics and analytics*, *2*, 9.

Yin, Y., Gao, J., Jones, B. F., & Wang, D. (2021). Coevolution of policy and science during the pandemic. *Science*, *371*(6525), 128-130.

Yu, H., Murat, B., Li, J., & Li, L. (2023). How can policy document mentions to scholarly papers be interpreted? An analysis of the underlying mentioning process. *Scientometrics*, *128*(11), 6247-6266.

https://doi.org/10.51408/issi2025_061

An Unsustainable Equation: Average Article Processing Charges Exceed Swedish Average PhD Salaries

A. I. M. Jakaria Rahman[1], Marco Schirone[2], Patrik Bergvall[3]

[1] *jakaria.rahman@chalmers.se*
Chalmers University of Technology, Hörsalsvägen 2, 41296 Gothenburg (Sweden)

[2]marco.schirone@chalmers.se
Chalmers University of Technology, Hörsalsvägen 2, 41296 Gothenburg (Sweden)
University of Borås, Allégatan 1, 50332 Borås (Sweden)

[3]*patrik.bergvall@chalmers.se*
Chalmers University of Technology, Hörsalsvägen 2, 41296 Gothenburg (Sweden)

Abstract

Open Access (OA) publishing has transformed scholarly communication by enhancing the visibility and accessibility of research. However, the rising costs of Article Processing Charges (APCs) pose significant financial challenges for researchers and institutions. In this paper, we investigated APC expenditure trends for publications from Swedish institutions, examining the relationship between total costs and publication volumes, variations in APCs among publishers, and the financial impact of gold and hybrid OA models over five years, focusing on six major academic publishers. Additionally, we explored disciplinary differences in APCs and access preferences, particularly between STEM (Science, Technology, Engineering, and Mathematics) and non-STEM fields. We sourced the publication dataset for this study from Scopus, including articles and reviews authored by researchers affiliated with Swedish institutions between 2019 and 2023. We categorized the publications using the SciVal tool and applied the Fields of Research and Development classification scheme to ensure structured and comparable disciplinary analysis. We obtained APC data from an openly available dataset and performed the analysis using a custom R script. Our findings reveal that OA publishing peaked in 2021, followed by a gradual decline, a trend likely driven by the surge in research dissemination during the COVID-19 pandemic. Total APC expenditure increased by 83%, rising from $12 million in 2019 to $22 million in 2023. Notably, the average APC exceeds the monthly average wage of Swedish PhD students, highlighting the financial burden of OA publishing. Hybrid OA models were found to be approximately 24% more expensive than gold OA models. Significant cost disparities were also observed among publishers. STEM fields incurred higher APCs than non-STEM fields, and a lack of gold OA journals in the Humanities was evident for several publishers. These findings highlight the financial strain associated with OA publishing and its uneven impact across disciplines and publishers. The study provides insights for policymakers, funding agencies, and academic institutions seeking to foster equitable and sustainable OA practices.

Introduction

The transition to Open Access (OA) publishing represents a transformative shift in academic publishing, fundamentally altering how research is disseminated, accessed, and funded. By removing paywalls, OA enhances the accessibility of scholarly work, increasing its visibility and fostering a wider dissemination across academic and non-academic audiences (Mikki, 2017; Tennant et al., 2016). OA also promotes transparency, reproducibility, and equitable access to scientific knowledge, fostering a more inclusive academic environment (Huang et al., 2024). Despite these benefits,

this transition is not without challenges. There are still gaps in understanding the economic implications of APCs across different publishing models, publishers, and disciplinary domains. A key issue is the rising cost of APCs, which are often required to publish in OA journals. A primary concern among researchers and institutions is the financial burden associated with OA publishing (Kendall, 2024; Segado-Boj et al., 2022). APCs required by many OA journals often put strain on institutional budgets, which raises questions about the sustainability of this model, especially for smaller universities and underfunded researchers (Borrego, 2023; Butler et al., 2023). These costs can place a heavy burden on researchers, institutions, and funding agencies, raising concerns about the long-term sustainability of OA publishing models (Asai, 2020; Shu & Larivière, 2024). This issue is particularly pronounced in the case of hybrid OA journals, which combine subscription-based access with an optional OA publishing route (Olsson, Lindelöw, et al., 2020). These financial pressures risk intensifying inequalities in the global research community, as authors from less funded institutions or regions may struggle to afford OA publication costs (Klebel & Ross-Hellauer, 2023). Several studies have noted the rising costs of APCs (Morrison, 2018; Pavan & Barbosa, 2018), raising concerns about the financial burden on researchers and institutions, particularly those from underfunded disciplines (Adegbilero-Iwari, 2024). These financial pressures have also been linked to growing disparities in access to OA publishing opportunities, especially for early-career researchers and non-STEM (Science, Technology, Engineering, and Mathematics) fields with limited funding (Nicholas et al., 2024).

In Sweden, OA publishing has grown significantly in the last decade, which is strongly supported by national policies, government directives, and mandates from research councils and funding agencies (SUHF, 2023). The backing provided by these initiatives puts an emphasis on the importance of open science and the principle that publicly funded research should be freely accessible to all. The Swedish Research Council, in collaboration with other key funding agencies such as Forte, Formas, and Vinnova, has mandated that research results must be openly accessible, emphasizing the principle that publicly funded research should benefit society at large (Swedish Research Council, 2022). This policy aligns with a broader commitment to ensure that publications appear exclusively in fully OA journals, enhancing the visibility and reach of Swedish research. The growing emphasis on OA in the Swedish academic landscape reflects both global trends and local priorities. As a result, OA has become a keystone of Sweden's research infrastructure, with universities and institutions actively promoting OA publishing models. Sweden provides a unique context for examining OA publishing challenges, given its strong commitment to OA and its well-established funding mechanisms for academic research. Despite these efforts, the high costs associated with OA publishing have become a growing concern (Frank et al., 2023).

However, comprehensive analyses of APC trends, their relationship to publication volumes, and cost disparities across major publishers and OA models remain unexplored in the Swedish context. Additionally, the variation in APCs between gold and hybrid OA models and among disciplinary domains, particularly between STEM and non-STEM fields, has not received attention. These gaps hinder the development

of reasonable and sustainable OA publishing frameworks, particularly in countries like Sweden, where national policies emphasize open science and publicly funded research mandates. Hence, we investigated the following research questions:

RQ1: How have APCs and publication volumes changed during a five-year period?
RQ2: How are total costs related to the number of publications during this period?
RQ3: How do APCs differ among six major publishers?
RQ4: How do APCs differ between gold and hybrid open access publishing models?
RQ5: How do APCs differ across disciplinary domains, particularly STEM versus non-STEM fields?

By addressing the above-mentioned research questions, we investigate trends and patterns in APCs for publications affiliated with Swedish institutions over a five-year period (2019–2023), focusing on six major academic publishers. Our analysis examines the financial dynamics of OA publishing, comparing the average costs associated with gold OA and hybrid OA models. We also explore disciplinary differences in APCs and the availability of gold OA and hybrid OA options, shedding light on the complex interplay between publishing costs, access, and academic disciplines. In this context, our goal is to provide empirical insights into the dynamics of APCs and offer evidence-based guidance to policymakers, funding agencies, and academic institutions for developing publication strategies that ensure the financial sustainability and inclusiveness of OA publishing.

Data and Methodology

Data for this study were retrieved from the Scopus database (Elsevier, 2025b) consisting of the metadata information of all articles and reviews authored by researchers affiliated with Swedish institutions between 2019 and 2023. Our dataset of Sweden-affiliated publications included 85,593 documents, of which approximately 71% (60,485) were identified as either gold OA or hybrid OA publications (see Table 1). Gold OA refers to publications that are freely available under an OA license, often accompanied by upfront APCs, while hybrid OA includes articles from subscription-based journals made OA through the payment of APCs. The dataset was cleaned to harmonize publisher names. For instance, Springer Nature, Springer, and Springer Science and Business Media Deutschland GmbH were all unified under the single name Springer. A similar standardization process was applied to other publishers.

Publications categorized in Scopus as ‘hybrid gold OA’ were treated exclusively as hybrid OA. If a publication was assigned multiple access types, such as ‘green OA; hybrid gold open,’ we classified it as hybrid OA. For cases where access types included combinations like ‘bronze OA; green OA,’ we retained both classifications as ‘bronze or green OA.’ Gold OA publications were kept unchanged in their original classification.

Table 1 presents the distribution of publications by six publishers categorized by access type—gold OA, hybrid OA, bronze or green OA and non-OA—along with a grand total for each publisher and access types. This categorization enables us to see the differences between open and non-OA trends among the publishers. Elsevier and Springer have relatively smaller shares of gold OA, reflecting their primary reliance on hybrid OA. In contrast, MDPI and Frontiers hold the largest shares of gold OA articles, as these publishers primarily operate under the gold OA model. Gold OA and hybrid OA account for 36% and 35% of the total publications in the table, respectively. Together, Swedish researchers published approximately 81% of their works as OA with these six publishers between 2019 and 2023. In this paper, we considered only the gold OA and hybrid OA publications to investigate the research questions.

Table 1. Distribution of publications between publishers and access types (2019 – 2023).

Publishers	*Gold Open Access*	*Hybrid Open Access*	*Bronze or Green Open Access*	*Non-Open Access*	*Grand Total*	
Wiley	2,212	7,266	3,145	2,470	15,093	(18%)
Springer	1,967	9,717	865	3,182	15,731	(18%)
Elsevier	4,689	12,953	5,207	10,239	33,088	(39%)
Frontiers	6,324	-	-	-	6,324	(7%)
MDPI	12,944	-	-	-	12,944	(15%)
PLoS	2,413	-	-	-	2,413	(3%)
Grand Total	30,549 (36%)	29,936 (35%)	9,217 (11%)	15,891 (18%)	85,593	

We used the SciVal tool (Elsevier, 2025a) to classify publications based on the major Fields of Research and Development (FORD) classification, as recommended by the Organization for Economic Co-operation and Development (OECD, 2015). This subject classification ensures consistency in grouping publications into relevant subject categories: Agricultural Sciences Engineering and Technology, Humanities, Medical Sciences, Natural Sciences, and Social Sciences, allowing for a more detailed understanding of APC variations across disciplines.

We obtained APC data from a publicly available dataset by Butler et al. (2024b), which provides APC values across six major publishers. To utilize the information conveyed by this dataset, we considered the same six publishers: Elsevier, Frontiers, MDPI, PLoS, Springer, and Wiley (see Table 1). The dataset reported the cost of APCs in US dollars and covered the same five-year period as the publication data considered in our study. Moreover, we utilized ISSN as a base for identifying the journal and corresponding publishers and matched the ISSN with the APC data for any kind of calculations done in this paper. This step was vital for accurate

comparisons. We conducted the analysis using the R programming language (R Core Team, 2024) for data processing, statistical analysis, and visualization.

It is challenging to investigate the costs of individual journals due to variations in pricing practices and the lack of transparency in bundled subscription models (Björk & Solomon, 2015). To address these difficulties, we used list prices for our analysis. These are publicly stated baseline prices set by publishers, often used as a standard reference point for pricing comparisons and analysis, as they provide a more standardized and comparable benchmark across publishers and are important components of market dynamics (for a recent game-theoretical discussion on this topic, see Haan et al., 2023). Ultimately, our approach, which focuses on analyzing APCs using list prices and excludes discounts and other negotiations, illustrates the projected maximum burden faced by Swedish universities when covering APCs.

Earlier studies have explored various aspects of APCs and their implications. For example, Solomon & Björk (2016) examined APC expenditures by universities in the USA and Canada, using the Web of Science (WoS) as the basis for publication data and employing subject mapping between Scopus and WoS. Butler et al., (2023) focused on APC revenues generated by six major publishers for gold and hybrid journals, also using WoS for publication data. Similarly, Pavan & Barbosa (2018) explored the economic sustainability of scientific journals that publish in OA. They collected APC data from the Directory of Open Access Journals and publishers' websites, classifying Brazilian-affiliated publications based on WoS subject categories. The publications were organized into specific subject areas and one multidisciplinary category. In our study, we retrieved Swedish-affiliated publications from Scopus and categorized them using the FORD classification, while incorporating tested APC data from Butler et al. (2024b). The FORD classification provides a high level of granularity, allowing for precise categorization of research outputs. Furthermore, it is often aligned with national research priorities and funding policies, making it a suitable framework for our analysis. This methodological approach contributes to the study of APCs by utilizing data available in Scopus and the categorization offered by the FORD classification.

Results

We investigated APCs and publication volumes focusing on trends and patterns. We assessed whether APCs have grown, plateaued, or fluctuated, and how these changes relate to the rise in publications in OA journals. Figure 1 presents a comparative analysis of the total APCs incurred and the number of publications produced annually during the period 2019–2023. The figure illustrates trends in APC expenditures alongside publication outputs, highlighting any correlations or disparities between the two variables over time. This data provides insights into the financial investments associated with OA publishing and the resulting research outputs, offering a basis for evaluating the cost-effectiveness and sustainability of the current publishing practice in Sweden.

We found significant changes in APCs from 2019 to 2023, with an 83% increase from $12 million in 2019 to $22 million in 2023 (see Figure 1). The most notable surge occurred between 2020 and 2021, with a 40% increase from $15 million to $21

million, primarily attributed to the implementation of transformative agreements (Widding, 2024) that converted traditional subscription costs to OA fees which is in line with the findings of Borrego et al., (2021) and Olsson et al., (2020). We found that OA publishing peaked in 2021 (8.4 thousand), followed by a moderate decline in 2022 (8.3 thousand), and a further decrease in 2023 (7.7 thousand). This pattern was significantly influenced by the global COVID-19 pandemic response, which indicates rapid research dissemination (Kim & Atteraya, 2023; Nane et al., 2023).

The above-mentioned findings suggest that the scholarly publishing landscape experienced a substantial transformation, driven by both institutional policy changes and extraordinary global circumstances. These results have significant implications for research funding allocation, institutional budgeting, and the future sustainability of OA publishing models. The observed trends highlight the need for continued monitoring of publishing costs and careful consideration of funding mechanisms for scholarly communication.

We examined the relationship between total costs and the number of publications, analyzing how variations in publication volume impact overall expenditure on APCs. A Pearson correlation analysis revealed a strong positive correlation between total costs and number of publications ($r = 0.85$, $p = 0.03$). The correlation coefficient indicates that as total costs increase, the number of publications tends to increase as well, with approximately 72% of the variance shared between these variables ($r^2 = 0.72$).

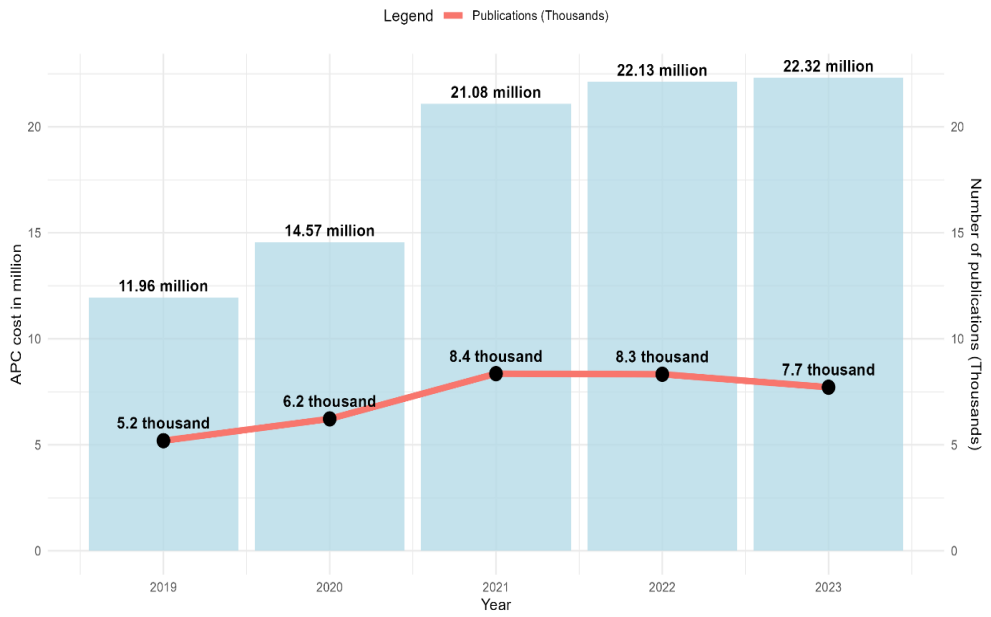


Figure 1. Total APCs and number of publications during 2019-2023.

The relationship was found to be statistically significant at the 0.05 level, suggesting this association is unlikely to have occurred by chance. Therefore, it can be predicted that 72% of the variation in the number of publications is attributable to APC costs, while the remaining 28% is influenced by other factors, such as research quality,

efficiency, or access to additional resources that are not directly related to cost (Björk & Solomon, 2015; Rowley et al., 2017; Xu et al., 2023). These results show the complex nature of publication dynamics and stress the importance of considering both financial and non-financial factors in academic output.

Furthermore, we examined the variations in APCs across six major publishers. Figure 2 illustrates the total APCs paid to six major publishers from 2019 to 2023. The data highlights trends in APC expenditure for each publisher over the five-year period, showcasing variations in costs and identifying patterns in publisher-specific spending. According to Figure 2, Elsevier dominates APC expenditure, reaching $38 million, which accounts for 41% of the total APC market. This significant financial dominance emphasizes Elsevier's established position as a key player in the scholarly publishing landscape. MDPI, with $23 million (24%), and Frontiers, with $17 million (18%), exhibit consistent and notable increases in APC costs, indicating their rapid market expansion and growing influence in the OA publishing sector. Similarly, Wiley ($8 million; 9%), Springer ($4 million; 5%), and PLoS ($3 million; 4%) are emerging as notable competitors, reflecting their strategic investment in OA publishing models. These findings highlight the evolving dynamics of the APC market, where Elsevier continues to maintain its dominance, while MDPI and Frontiers solidify their positions as key challengers.

Meanwhile, Wiley, Springer, and PLoS are gradually increasing their presence, highlighting a diversified growth across different publishers. This trend aligns with previous studies, suggesting a competitive shift in the global scholarly communication market as publishers adapt to the growing demand for OA (Borrego, 2023; Halevi et al., 2024). These findings provide insights into how the APC market is shaping the broader academic publishing ecosystem.

Figure 3 illustrates the distribution of APCs across six major publishers, comparing costs between gold and hybrid OA publishing models. The figure highlights the average APC (the dash line) for each publisher within these two categories, providing a clear visualization of cost disparities.

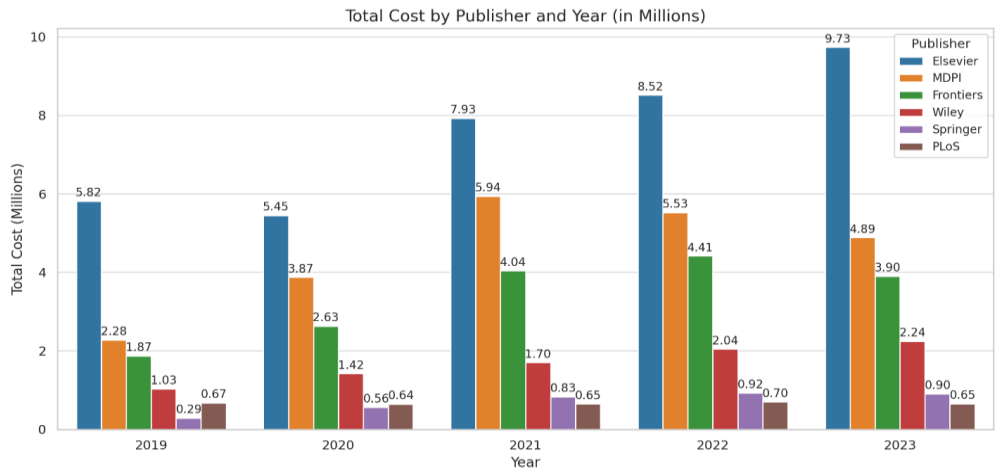


Figure 2. Total APCs by six major publishers from 2019 to 2023.

Furthermore, we examined the differences in APCs between gold and hybrid OA models and their average publishing costs. Notably, hybrid OA models consistently exhibit higher APCs compared to gold OA models across most publishers (Mittermaier, 2015). The data emphasizes significant variation in APCs among publishers, suggesting potential differences in pricing strategies. We found that APCs for gold OA range from $1,750 to $3,100, with an average cost of $2,900, offering relatively lower and more variable pricing. In contrast, hybrid OA is characterized by consistently higher costs, with APCs ranging from $2,600 to $4,950 and an average of $3,800. This makes hybrid OA approximately $900 (24%) more expensive per article than gold OA. This cost disparity reflects established trends in the scholarly publishing industry, where hybrid journals charge significantly higher APCs compared to fully OA journals.

Early-career researchers, such as PhD students, often face significant financial barriers, as the average cost of publishing a single article in a gold ($2,900,) or hybrid OA ($3,800) journal even exceed the average monthly salary of a PhD student in Sweden which is around $2,850 (SCB, 2023). PhD students are often affiliated with universities or funded through grants and scholarships. However, this financial support does not always cover APCs. University scholarships or doctoral funding schemes primarily support living expenses, tuition, and research activities, rather than publication costs. APCs are frequently excluded from standard research budgets unless specifically requested or allocated in advance (Wang, 2024). Competitive research grants that cover APCs are typically awarded to senior researchers or principal investigators, leaving PhD students to navigate the publication process with limited financial autonomy. Even when institutional OA agreements or funds exist, they may only apply to selected journals or be subject to annual caps, making access inconsistent. Consequently, early-career researchers may find it challenging to publish in reputable gold or hybrid OA journals, despite producing high-quality research (Nicholas et al., 2024).

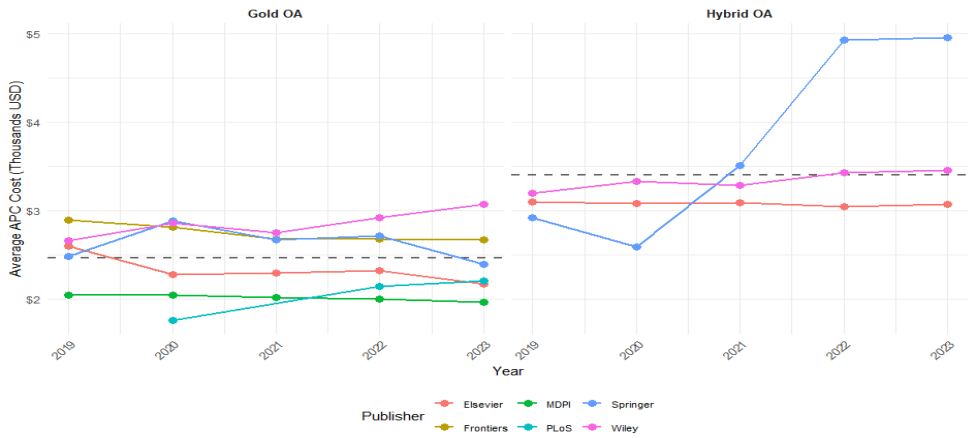


Figure 3. APCs by publishers in Gold OA and Hybrid OA with their average cost.

In this context, we argue that no researcher should have to allocate the equivalent of an entire month's salary of a PhD student just to publish their work openly. While

funding mechanisms exist, the current pricing models of OA publishing challenge the fundamental principle of equitable access and place excessive financial pressure on the very researchers that open science aims to empower. This finding raise questions about the accessibility and equity of current OA publishing models, not only for researchers but also for funding agencies and institutions tasked with supporting open scholarship (Khoo, 2019).

Moreover, the higher costs associated with hybrid OA have been criticized for contributing to the so-called "double-dipping" phenomenon, where publishers charge both subscription fees and APCs for hybrid journals, adding an additional financial burden to academic institutions (Asai, 2023b). Thus, we suggest greater scrutiny and transparency in APC pricing structures and advocate for the adoption of cost-effective and reasonable publishing practices, particularly as the global academic community shifts toward OA mandates and transformative agreements.

In addition, we investigated the differences in APCs across disciplinary domains, with a particular focus on comparing STEM and non-STEM fields. Figure 4 presents the average APCs for the different publishers, categorized by the FORD classification and gold and hybrid OA publishing model. The figure illustrates how APCs vary not only between publishers but also within specific disciplines, highlighting the disparities in publishing costs for different disciplines. It also compares the average APCs between gold and hybrid OA models, revealing whether certain fields or access types are more associated with higher publishing costs.

We found that across most subject areas, hybrid OA consistently incurs higher average APCs compared to gold OA, with notable exceptions in specific disciplines such as Engineering and Technologies. This trend is largely attributed to the traditional publishing models employed by major publishers, where hybrid journals often tend to impose higher APCs to cover both subscription and OA costs (Asai, 2023a).

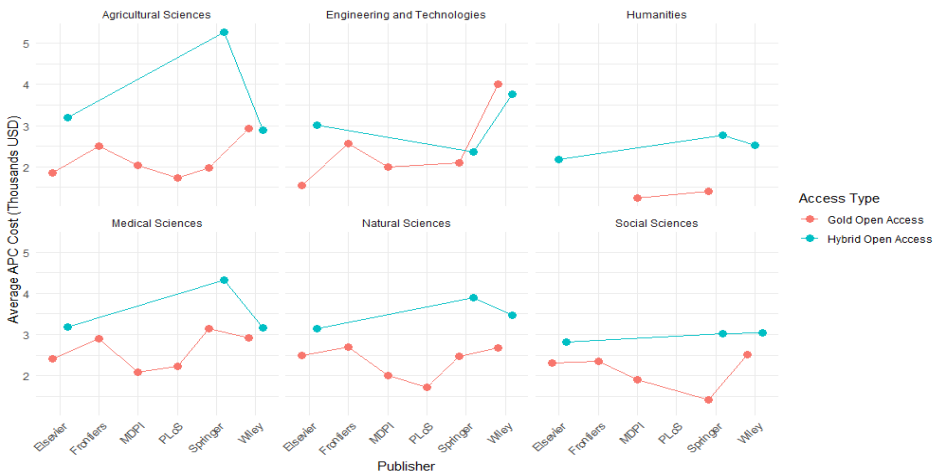


Figure 4. Average APCs by FORD classification by publisher and access types (2019-2023).

Further, we found Springers´ APCs for hybrid OA to be significantly higher across all subject areas, which emphasizes the association with hybrid publishing models. Interestingly, the availability of gold OA journals varies by discipline. For instance, in Humanities, gold OA options are limited, with publishers like PLoS and Springer being among the few that offer fully OA journals in this field. This limited availability can constrain researchers' options and influence their publishing decisions, particularly in fields where hybrid models dominate the OA landscape. Fields such as Agricultural Sciences, Medical Sciences, and Natural Sciences often incur higher APCs compared to disciplines like Social Sciences and Humanities (See Figure 4).

The above-mentioned disparities highlight the unequal financial burdens faced by researchers, which are shaped by publishing practices and market dynamics within their respective disciplines. The higher costs of hybrid OA, coupled with limited gold OA options in certain disciplines, pose challenges for researchers, especially those with constrained budgets or from underfunded institutions (Morillo, 2020; Perianes-Rodríguez & Olmeda-Gómez, 2021).

Such cost variations stress the importance of developing field-specific OA publication strategies to ensure equitable access to OA publishing opportunities. Furthermore, the differential pricing between gold and hybrid OA raises questions about the sustainability of the current publishing ecosystem. This calls for greater advocacy for affordable OA models, increased support for fully OA journals, and transparency in APC pricing to foster a more inclusive scholarly publishing environment.

Discussion

We investigated the evolving landscape of APCs and publication volumes in the context of OA publishing, focusing on Sweden's research output over a five-year period. Our findings address key research questions, providing a comprehensive understanding of the financial and disciplinary dynamics of OA publishing. In *RQ1*, we investigated the dynamics of APCs and publication volumes over five years to identify trends and shifts that could inform publishing practices. We found that between 2019 and 2023, APC expenditures grew by 83%, with the most significant increase of 40% occurring between 2020 and 2021 due to transformative agreements. OA publishing peaked in 2021, driven by the need for rapid research dissemination during the COVID-19 pandemic, before declining moderately in the following years. The findings demonstrate that transformative agreements have accelerated the transition to OA but also contributed to rising APC costs, highlighting the financial implications of such policies (Inchcoombe et al., 2022; Widmark, 2024).

To address *RQ2*, we analyzed the relationship between total APC costs incurred and the volume of publications during the observed period, aiming to uncover patterns in expenditure efficiency. Our analysis demonstrated a strong positive correlation between total APC costs and the number of publications, indicating that as APC costs increase, the number of publications also tends to rise. The correlation coefficient suggests that approximately 72% of the variance in publication volume can be explained by total APC costs. These findings highlight a direct and statistically

significant relationship between the financial investment in APCs and the increase in publication output, emphasizing the economic implications of OA publishing (Björk & Solomon, 2015).

Moreover, answering *RQ3*, we investigated APC variations among six major publishers to explore economic disparities across publishing platforms. Significant disparities in APCs were observed among the six major publishers. Elsevier leads the APC market, accounting for 41% and demonstrating its dominant role in scholarly publishing. MDPI (24%) and Frontiers (18%) are rapidly expanding, showing significant growth in market influence. Wiley (9%), Springer (5%), and PLoS (4%) are also emerging as notable competitors, reflecting a diversification of the APC market with increasing contributions. These differences highlight variations in publishers' pricing strategies and their implications for authors and institutions (Asai, 2020; Budzinski et al., 2020).

Furthermore, in *RQ4*, we evaluated the differences in APCs between gold and hybrid OA publishing models, providing insights into the financial implications of each model. We found significant disparities between gold and hybrid OA models. Gold OA journals typically charge lower APCs, averaging $2,900, whereas hybrid OA journals charge consistently higher fees, averaging $3,800—a 24% premium. This pricing structure particularly impacts early-career researchers, as both models exceed the average monthly salary of PhD students in Sweden ($2,850), highlighting significant barriers in academic publishing (Green, 2019; L. Zhang et al., 2022). Simultaneously, there are established industry practices in which traditional subscription-based publishers maintain dual revenue streams, which also affect institutional library budgets.

In *RQ5*, we analyzed APC variations across disciplines, emphasizing differences between STEM and non-STEM fields. We found that the availability and costs of gold OA journals vary significantly across disciplines, impacting researchers' publishing decisions. In fields like the Humanities, gold OA options are limited, with publishers such as Springer and MDPI offering some of the few fully OA journals. This scarcity contrasts with hybrid models that dominate the OA landscape in all disciplines. Further, disciplines like Agriculture, Medical Science, and Natural Sciences face higher APCs compared to fields like the Humanities, or the Social Sciences. These disparities create unequal financial burdens for discipline-specific researchers and institutions with limited budgets (Morillo, 2020; X. Zhang et al., 2020). The above findings emphasize the need for transparent pricing, reasonable funding mechanisms, and policies that support sustainable and inclusive OA publishing strategies across all disciplines.

The discussion above demonstrated that APCs expenditure increased by 83% during this period, with a sharp 40% rise between 2020 and 2021. This increase was largely due to transformative agreements and the surge in publishing during the COVID-19 pandemic. While these agreements accelerated OA adoption, they also contributed to rising costs, signaling financial sustainability concerns. A strong positive correlation between total APC costs and publication volume confirms that increased financial investment leads to higher output. However, this also emphasizes the need for more cost-efficient publishing strategies. Significant disparities were found

among publishers. Elsevier dominated the APC market, followed by MDPI and Frontiers. These variations reflect differing pricing models and market concentration, which influence authors' choices and institutional budgets. Additionally, hybrid OA journals charge a 24% premium compared to gold OA journals, making them a less affordable option. This dual-cost model of hybrid OA also strains institutional library funds. Disciplinary analysis revealed that researchers in STEM fields face higher APCs, while those in the Humanities and Social Sciences encounter limited gold OA options. This highlights unequal access and funding burdens across disciplines.

We argue that OA publishing has expanded in Sweden, but it is still affected by cost imbalances, the dominance of major publishers, and disparities in access across disciplines. To promote a more equitable and sustainable OA future, greater transparency in pricing, targeted funding support, and inclusive policy development are essential. To ensure a fair and sustainable OA ecosystem, it is imperative for policymakers to implement stricter regulations on APC pricing and to demand greater transparency and accountability from publishers benefiting from public funds. One reviewer pointed out that the multinational initiative 'cOAlition S' (Schiltz, 2018) has not succeeded in limiting APC costs and that the anticipated transformation of the scholarly publishing system has yet to materialize, motivating us to consider how we should respond to this concern. We argue that since 2018, 'cOAlition S' has promoted transformative agreements as a strategy to transition scholarly publishing toward immediate OA. However, several challenges have limited their ability to control APC pricing and to fully realize a systemic transformation (Brainard, 2024). In light of this, we suggest that the Swedish government could draw lessons from the experience of 'cOAlition S'. By engaging in strategic dialogue with the publishers most frequently used by Swedish researchers. Sweden should be able to develop a more targeted approach that ensures the best return on taxpayers' money.

Limitations and Future research

This study analyzed publications with at least one author affiliated with a Swedish higher education institution. While this does not confirm that the Swedish author(s) directly paid the APCs, it is reasonable to assume that they were associated with these costs, albeit to a varying degree. Factors such as agreements between authors, institutional policies on APC payments, discounts, waivers, and other variables contribute to the varying degrees of financial responsibility. As emphasized in the data and methodology section, this study focuses on estimating the projected maximum burden faced by Swedish universities when covering APCs. However, as noted by Butler et al., (2024a), APC data collection is inherently complex and may include gaps, meaning that not all journals in this study have corresponding APC information. This limitation highlights the challenges of comprehensively mapping APC trends across publishers and disciplines. Additionally, this study includes only six publishers, though many other legitimate publishers support OA publishing. Including other publishers would likely reveal significantly higher total expenses for OA publishing.

We aim to further study encompassing Nordic countries to gain a more comprehensive understanding of regional trends in OA publishing. Additionally, a comparative analysis between the actual costs incurred under transformative agreements with publishers and the listed APCs would provide insights for policymakers. Such an approach could help assess the economic implications of current agreements and inform future strategies for sustainable OA publishing by Sweden affiliated researchers.

Conclusions

Our findings highlight the increasing financial burden associated with OA publishing, particularly within Sweden, where transformative agreements and institutional policies are reshaping the publication landscape. While these transformative agreements have accelerated the adoption of OA models, they have also significantly elevated APC expenditures, showing financial consequences for such policies. The strong correlation between APC costs and publication volumes points out the economic trade-offs involved in achieving higher research output. The predominance of major publishers, the emergence of new players, and the persistent disparities in APCs across different models and disciplines emphasize areas for negotiation and policy development. Notably, the cost of OA publishing frequently surpasses the monthly salary of a PhD student in Sweden, a fact that necessitates attention from policymakers. This issue is especially concerning given that publishers receive substantial funding from taxpayer money, yet there is an oversight regarding the pricing of OA publishing. The absence of standardized pricing mechanisms or accountability for the use of public funds enables publishers to set APCs arbitrarily, thereby creating financial barriers for researchers and underfunded institutions. By mapping APC trends and identifying key cost drivers, this study provides insights for policymakers, institutions, and researchers to promote more equitable and sustainable OA practices. Future research could expand its scope to encompass broader geographic regions or analyze the long-term impacts of transformative agreements on publication costs and accessibility.

Acknowledgements

We express our sincere gratitude to the anonymous reviewers of the *20th International Conference on Scientometrics and Informetrics* (ISSI, 2025), Yerevan, Armenia, for their insightful feedback.

Disclaimer

We have presented an initial version of this study at the *29th Nordic Workshop on Bibliometrics & Research Policy*, November 20-22, 2024, in Reykjavík, Iceland. This research received no external funding or financial support.

References

Adegbilero-Iwari, I. (2024). From serials crisis to dollar crisis: the compelling evidence against APC-based open access in sub-Saharan Africa countries. *Learned Publishing*.

Asai, S. (2020). Market power of publishers in setting article processing charges for open

access journals. *Scientometrics*, *123*(2), 1037–1049.
Asai, S. (2023a). Determinants of article processing charges for hybrid and gold open access journals. *Information Discovery and Delivery*, *51*(2), 121–129.
Asai, S. (2023b). Does double dipping occur? The case of Wiley's hybrid journals. *Scientometrics*, *128*(9), 5159–5168.
Björk, B. C., & Solomon, D. (2015). Article processing charges in OA journals: relationship between price and quality. *Scientometrics*, *103*(2), 373–385.
Borrego, Á. (2023). Article processing charges for open access journal publishing: a review. *Learned Publishing*, *36*(3), 359–378.
Borrego, Á., Anglada, L., & Abadal, E. (2021). Transformative agreements: do they pave the way to open access? *Learned Publishing*, *34*(2), 216–232.
Brainard, J. (2024). *A mixed review for Plan S's drive to make papers open access: evaluation describes unintended effects as funders mull expanding the policy*. Science. https://doi.org/10.1126/SCIENCE.Z6IZNOC
Budzinski, O., Grebel, T., Wolling, J., & Zhang, X. (2020). Drivers of article processing charges in open access. *Scientometrics*, *124*(3), 2185–2206.
Butler, L.-A., Hare, M., Schönfelder, N., Schares, E., Alperin, J. P., & Haustein, S. (2024a). An open dataset of article processing charges from six large scholarly publishers (2019-2023). *ArXiv*. https://arxiv.org/pdf/2406.08356
Butler, L.-A., Hare, M., Schönfelder, N., Schares, E., Alperin, J. P., & Haustein, S. (2024b). Open dataset of annual Article Processing Charges (APCs) of gold and hybrid journals published by Elsevier, Frontiers, MDPI, PLOS, Springer-Nature and Wiley 2019-2023. *Harvard Dataverse*. https://doi.org/10.7910/DVN/CR1MMV
Butler, L.-A., Matthias, L., Simard, M. A., Mongeon, P., & Haustein, S. (2023). The oligopoly's shift to open access: how the big five academic publishers profit from article processing charges. *Quantitative Science Studies*, *4*(4), 778–799.
Elsevier. (2025a). *SciVal*. Retrieved from www.scival.com.
Elsevier. (2025b). *Scopus*. Retrieved from www.scopus.com.
Frank, J., Foster, R., & Pagliari, C. (2023). Open access publishing – noble intention, flawed reality. *Social Science and Medicine*, *317*.
Green, T. (2019). Is open access affordable? Why current models do not work and why we need internet-era transformation of scholarly communications. *Learned Publishing*, *32*(1), 13–25.
Haan, M. A., Heijnen, P., & Obradovits, M. (2023). Competition with list prices. *Games and Economic Behavior*, *140*, 502–528.
Halevi, G., Jiménez, R. S., Bote, V. P. G., & Anegón, F. D.-M. (2024). Estimating the financial value of scientific journals and APCs using visibility factors: a new methodological approach. *Profesional de La Información*, *33*(5).
Huang, C. K., Neylon, C., Montgomery, L., Hosking, R., Diprose, J. P., Handcock, R. N., & Wilson, K. (2024). Open access research outputs receive more diverse citations. *Scientometrics*, *129*(2), 825–845.
Inchcoombe, S., Winter, S., Lucraft, M., & Baker, K. (2022). Transforming Transformative Agreements. *Logos*, *32*(4), 7–14.
Kendall, G. (2024). Are open access fees a good use of taxpayers' money? *Quantitative Science Studies*, *5*(1), 264–270.
Khoo, S. Y. S. (2019). Article processing charge hyperinflation and price insensitivity: an Open Access sequel to the serials crisis. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *29*(1), 1–18.
Kim, E., & Atteraya, M. S. (2023). A decade of changes in OA and non-OA journal

publication and production. *Journal of Librarianship and Information Science*.
Klebel, T., & Ross-Hellauer, T. (2023). The APC-barrier and its effect on stratification in open access publishing. *Quantitative Science Studies*, *4*(1), 22–43.
Mikki, S. (2017). Scholarly publications beyond pay-walls: increased citation advantage for open publishing. *Scientometrics*, *113*(3), 1529–1538.
Mittermaier, B. (2015). Double dipping in hybrid open access – chimera or reality? *ScienceOpen Research*, *0*(0).
Morillo, F. (2020). Is open access publication useful for all research fields? Presence of funding, collaboration and impact. *Scientometrics*, *125*(1), 689–716.
Morrison, H. (2018). Global OA APCs (APC) 2010–2017: major trends. *Electronic Publishing*, *339*(88).
Nane, G. F., Robinson-Garcia, N., van Schalkwyk, F., & Torres-Salinas, D. (2023). COVID-19 and the scientific publishing system: growth, open access and scientific fields. *Scientometrics*, *128*(1), 345–362.
Nicholas, D., Revez, J., Abrizah, A., Rodríguez-Bravo, B., Boukacem-Zeghmouri, C., Clark, D., Xu, J., Swigon, M., Watkinson, A., Jamali, H. R., & Herman, E. (2024). Purchase and publish: early career researchers and open access publishing costs. *Learned Publishing*, *37*(4), e1617.
OECD. (2015). Concepts and definitions for identifying R&D. In *Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development*.
https://www.oecd.org/en/publications/frascati-manual-2015_9789264239012-en.html
Olsson, L., Francke, H., Lindelöw, C. H., & Willén, N. (2020). The first Swedish read & publish agreement: an evaluation of the springer compact pilot. *LIBER Quarterly*, *30*(1), 1–33.
Olsson, L., Lindelöw, C. H., Österlund, L., & Jakobsson, F. (2020). Cancelling with the world's largest scholarly publisher: lessons from the Swedish experience of having no access to Elsevier. *Insights: The UKSG Journal*, *33*.
Pavan, C., & Barbosa, M. C. (2018). Article processing charge (APC) for publishing open access articles: the Brazilian scenario. *Scientometrics*, *117*(2), 805–823.
Perianes-Rodríguez, A., & Olmeda-Gómez, C. (2021). Effect of policies promoting open access in the scientific ecosystem: case study of ERC grantee publication practice. *Scientometrics*, *126*(8), 6825–6836.
R Core Team. (2024). *R: a language and environment for statistical computing*. https://www.r-project.org
Rowley, J., Johnson, F., Sbaffi, L., Frass, W., & Devine, E. (2017). Academics' behaviors and attitudes towards open access publishing in scholarly journals. *Journal of the Association for Information Science and Technology*, *68*(5), 1201–1211.
SCB. (2023). *Average monthly salary and salarydispersion by occupation (SSYK) and sex, 2023*. Statistics Sweden.
Segado-Boj, F., Prieto-Gutiérrez, J.-J., Martín-Quevedo, J., Segado-Boj, F., Prieto-Gutiérrez, J.-J., & Martín-Quevedo, J. (2022). Attitudes, willingness, and resources to cover article publishing charges: the influence of age, position, income level country, discipline and open access habits. *Learned Publishing*, *35*(4), 489–498.
Shu, F., & Larivière, V. (2024). The oligopoly of open access publishing. *Scientometrics*, *129*(1), 519–536.
Solomon, D., & Björk, B. C. (2016). Article processing charges for open access publication-the situation for research intensive universities in the USA and Canada. *PeerJ*, *2016*(7), e2264. https://doi.org/10.7717/PEERJ.2264/SUPP-1

SUHF. (2023). *Recommendation regarding charting Sweden's path beyond the transformative agreements*. https://suhf.se/app/uploads/2024/01/SUHF-REC-2023-7-Recommendation-regarding-charting-Swedens-path-beyond-the-transformative-agreements.pdf

Swedish Research Council. (2022). *Guidelines for publishing with open access - Swedish Research Council*. https://www.vr.se/english/mandates/open-science/open-access-to-scientific-publications/guidelines-for-publishing-with-open-access.html

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., & Hartgerink, C. H. J. (2016). The academic, economic and societal impacts of Open Access: an evidence-based review. *F1000Research*, *5*, 632.

Wang, J. (2024). Article processing charges suppress the scholarship of doctoral students. *European Science Editing*, *50*.

Widding, A. S. (2024). Beyond transformative agreements: ways forward for universities. *European Review*, *32*(S1), S28–S38.

Widmark, W. (2024). How can we get beyond the transformative agreements: a Swedish perspective. *Revista Española de Documentación Científica*, *47*(4).

Xu, X., Xie, J., Sun, J., & Cheng, Y. (2023). Factors affecting authors' manuscript submission behaviour: a systematic review. *Learned Publishing*, *36*(2), 285–298.

Zhang, L., Wei, Y., Huang, Y., & Sivertsen, G. (2022). Should open access lead to closed research? The trends towards paying to perform research. *Scientometrics*, *127*(12), 7653–7679.

Zhang, X., Grebel, T., & Budzinski, O. (2020). *The prices of open access publishing: the composition of APC across different fields of sciences*.

https://doi.org/10.51408/issi2025_070

Analysis of Relationships Between Paper Citations and Their Category Influencing Factors: A Bayesian Network with Latent Variables Approach

Mingyue Sun[1], Mingliang Yue [2], Wen Peng[3], Tingcan Ma[4]

[1]*sunmingyue22@mails.ucas.ac.cn*, [2]*yueml@mail.whlib.ac.cn*, [3]*pengwen23@mails.ucas.ac.cn*, [4]*matc@whlib.ac.cn*

Chinese Academy of Sciences, National Science Library (Wuhan), 430071 Wuhan (China)
University of Chinese Academy of Sciences, Department of Information Resources Management, School of Economics and Management, 100190 Beijing (China)

Abstract

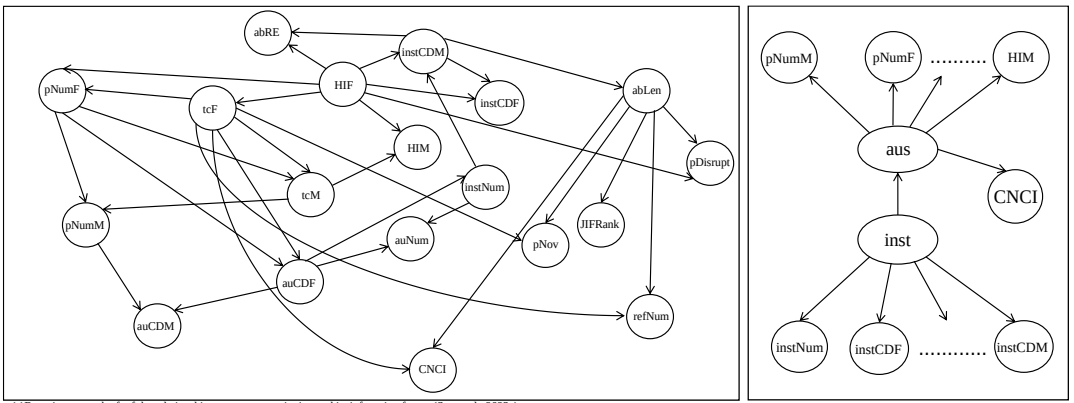
The analysis of the impact of academic papers has long been a topic of interest among scholars. Many studies have been carried out to explore the interaction between paper citations and its influencing factors from a microscopic perspective, e.g., analyzing the correlation between individual or multiple observable variables (such as author h index and publication counts) and citation count. However, it remains challenging to conduct analysis from a relatively macroscopic perspective, such as understanding how author characteristics as a whole influence citation count. In this paper, we adopt a Bayesian Network (BN) with latent variables as the knowledge framework, using latent variables to describe characteristics of different aspects (i.e., institution, author and paper aspects) as a whole, so that interactions among latent category factors as well as observable factors can be analyzed. We use the K-means algorithm to acquire categories of latent variables and use constraint-based scoring approach to learn the BN. We analyzed how the introduction of latent variables provides new perspectives compared to using only observable variables, conducted corresponding analyses, and reached certain conclusions.

Introduction

Citation plays a crucial role in the scientific evaluation of publications, individual scientists, and research institutions, prompting the academic community to contemplate the mechanisms and rationale behind its use for evaluation purposes. Numerous scholars have studied the factors that influence citation rates and how they affect the number of citations (Bornmann, 2011; Xie et al., 2019).

According to Tahamtan and Bornmann (2018a), the process of a research paper being cited is complex. There are significant relationships between paper citations and various factors (Xie et al., 2019), including authorship characteristics such as academic influence, gender, academic background, and others (Hurley et al., 2013; Ruan et al., 2020; Stremersch et al., 2015; Wang et al., 2019a, 2019b), as well as institutional and/or national affiliations (Didegah & Thelwall, 2013). Other influential factors include the impact of the publishing journal (Bornmann & Leydesdorff, 2015; Stegehuis et al., 2015), linguistic properties of the paper such as readability (Didegah & Thelwall, 2013; Stremersch et al., 2015), the paper's innovativeness (Wu et al., 2019), the number and impact of references (Bornmann & Leydesdorff, 2015), and other considerations like scientific funding (Rigby, 2013) and open access status (McCabe & Snyder, 2014), among others.

While previous studies have analyzed the independent or joint associations between various factors and paper citations, there has been relatively insufficient consideration of the correlations between these influencing factors. Sun et al. (2023) addressed this gap by applying Bayesian network (BN) to study the interactive relationships between citation and its influencing factors, utilizing 20 variables. However, the constructed network structure may be complex. As depicted in Figure 1(a), the intricate dependency relationships represented by directed edges between nodes may hinder effective focus on specific analyses of interest, such as understanding how author factors as a whole influence the citation impact of papers. To simplify the BN structure and facilitate more intuitive inference, latent variables can be introduced (Koller & Friedman 2009, Zhang & Guo 2006). For instance, the introduction of latent variables, represented as *aus* (author factors) and *inst* (institutional factors), significantly streamlined the dependency relationships between variables, as shown in Figure 1(b). Meanwhile, as demonstrated in the Results section, this streamlined network can allows for new analytical perspectives.



(a)Bayesian network of of the relationships among paper citation and its infuencing factors(Sun et al., 2023a).

(b)Bayesian network of relationships between author factors and institutional factors with two latent variables.

Figure 1. Bayesian Network (with latent variables).

Therefore, this paper adopts a Bayesian network incorporating latent variables (as categorical factors) influencing paper citations, including author factor, institutional factor, and factor related to paper-specific characteristics. To differentiate from traditional BN models that directly use observable variables (Sun et al., 2023) , this study applies a domain-specific latent variable learning method. This method captures implicit patterns across multiple observable variables, enabling the BN structure to reflect higher-level macroscopic interrelations between latent variables with reduced complexity.The BN structure is learned based on a constraint-based scoring algorithm that incorporates domain expert knowledge. After modeling, BN inferences are conducted to discover new analytical perspectives and draw conclusions.

The remainder of this paper is structured as follows: Section 2 introduces the necessary knowledge of Bayesian networks with latent variables. Section 3 outlines the construction process of the BN, including optimal structure learning and

parameter learning of BN with latent variables. Section 4 demonstrates BN inference and presents findings. Section 5 concludes the paper.

Bayesian network with latent variables

A Bayesian network with latent variables (Zhang & Guo, 2006) is defined as a binary tuple (G, θ). $G = (\chi, E)$represents a DAG structure, where the node set $\chi = \{x_1, \ldots, x_n\}$ consists of each node corresponding to a random variable, $\chi = O \cup \mathcal{L}$ indicates that χ includes both the observable variable set O and the latent variable set $\mathcal{L}$, with $|O| + |\mathcal{L}| = n$、$|O| > 0\ and\ |\mathcal{L}| \geq 0$. E represents the set of directed edges, where a directed edge $X_a \rightarrow X_b$ indicates a dependency relationship from node $X_{\square}$ to node $X_{\square}$, or a causal relationship where $X_{\square}$ is a direct cause of $X_{\square}$. θ represents the set of conditional probability parameters, denoted as $\pi(X_i) = \{X_j | < X_j, X_i > \in E\}$. If all nodes are discrete variables, $\theta_{\square} = \{P(X_i|\pi(X_i))\}$ represents the conditional probability distribution (CPT) of node $X_{\square}$, and $\theta_{\square\square\square} = \{P(X_i = k|\pi(X_i) = j)\}$ represents the conditional probability parameter corresponding to the situation where node $X_{\square}$ takes on value *k* and its parent nodes take on the *j*th combination of values. In Figure 1(b), the set of latent variables are {*aus*, *inst*}, and the set of observable variables is {*pNumM*, *pNumF*, *HIM*, *instCDM*, *instCDF*, *instNum*, ...}.

The construction of Bayesian network with latent variables mainly consists of four parts: determining the number of latent variables, determining the cardinality of latent variables, structure learning, and parameter learning (Koller & Friedman, 2009; Zhang & Guo, 2006), as demonstrated in Fig. 2. Determining the number of latent variables involves deciding how many latent variables are needed in the model. Methods for this include clustering-based approaches (Wang et al., 2008; Mourad et al., 2013) and clique-based methods (Elidan et al., 2000; He et al., 2014). Determining the cardinality of latent variables refers to determining the number of states each latent variable can take. Typically, clustering techniques are employed, treating latent variables as hidden categories. The process involves starting with a small number of categories (e.g., binary) and incrementally increasing them until the objective function reaches a maximum, with the corresponding category number considered as the cardinality of latent variables (Elidan & Friedman, 2013). In practice, the number and cardinality of latent variables are highly domain-specific and often determined by experts after analyzing the scenario (Wu & Yue, 2023). Structure learning aims to find the optimal network structure that fits the real data best using scoring-based search (Chickering, 2002; Ramsey et al., 2017; Goudet et al., 2018; Zhu et al., 2019) or conditional independence evaluation (Kong & Wang, 2023; Colombo & Maathuis, 2014). Parameter learning algorithms commonly utilize the EM algorithm and its variants (Qi et al., 2022; Kan et al., 2022).

In the network, latent variables often serve as abstractions of multiple observable variables, capturing the combined effects of the observable variables. Therefore, latent class model is usually adopted as local structure to model the relationships among latent variables and their corresponding observable variables. In the latent class model, observable variables are only connected to their corresponding latent class variables, and do not connect with other variables (Zhang & Guo, 2006). The

latent variables and other (latent or observable) variables can be interconnected to form a global network, thereby establishing relationships between the represented observable variables and other variables. The structure learning process is then used to determine the global structure, based on certain domain-specific constraints (Yue et al., 2020).

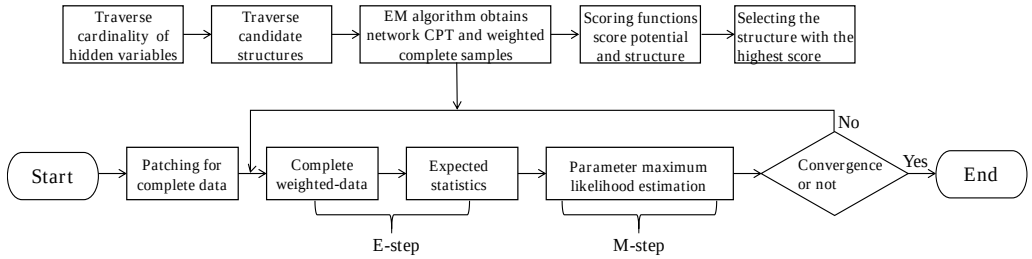


Figure 2. The flowchart for learning the structure and parameters of a Bayesian network with latent variables.

Bayesian Network construction

In this section, we first introduce the latent and observable variables considered in the BN. Then we discuss the constraints on the global structure based on the nature of academic citation. Finally, we present the BN construction algorithm.

The latent variables and corresponding observable variables

Unlike Sun et al., (2023) that focus on analyzing individual observable variables, this paper introduces three latent variables—paper factors, inst_factors and aus_factors—to abstract and integrate diverse observable variables into higher-level macroscopic dimensions. The rationale for selecting these latent variables is grounded in the citation mechanism outlined by Tahamtan and Bornmann (2018), which highlights the multifaceted influences on a paper's ability to garner citations. According to their findings, the intrinsic value of a research paper is a key determinant of its academic influence, while author characteristics significantly shape the citing author's expectations of the document's value. These author characteristics are further categorized into Author-level factors and Platform-level factors, both of which are posited to influence the perceived value of a paper. Building on this understanding, this study identifies three latent variables—paper factors, inst_factors and aus_factors.This latent variables simplifies the representation of complex observable variables interactions while preserving critical dependencies of paper itself, authors and institutes.

Paper_factors represents a latent variable that integrates multiple observable variables (characteristics) of the paper' s research content, which collectively capture the paper' s academic value. *Paper_factors* is categorized into four categories based on a comprehensive assessment of various observable variables, such as novelty (pnov) (Bu et al. 2021) and the number of references and the citations they received (refNum, refcitation_sum, refcitation_average) (Rigby 2013; Onodera and Yoshikane 2015; Xie et al. 2019; Bornmann and Leydesdorff 2015). Each

category represents a distinct level of academic value, reflecting the combined influence of these related observable variables. These related obesrvable variables encompasses the novelty (pnov) (Bu et al. 2021) and disruptiveness (pDisrupt) (Wu et al. 2019) of a paper, reflecting aspects of a research work's contribution. The number of references and the citations they received (refNum, refcitation_sum, refcitation_average) (Rigby 2013; Onodera and Yoshikane 2015; Xie et al. 2019; Bornmann and Leydesdorff 2015) reflect the amount of knowledge and impact of knowledge referenced by the work, as well as linguistic properties influencing other researchers' understanding of the paper. This includes text readability (abER) (Stremersch et al. 2015; Lei and Yan 2016; Ante 2022) and text length (abstract_length, title_length, key_words_length) (Vamclay 2013; Xie et al. 2019; Ruan et al. 2020; Stremersch et al. 2015).

aus_factors represents a latent variable that integrates multiple observable variables (characteristics) related to the impact of authors of the research papers. *aus_factors* is categorized into four categories based on a comprehensive assessment of various observable variables, such as the number of papers published (pNum_Max, pNum_average, pNumF) (Stremersch et al. 2015), the number of citations received by published papers (tc_Max, tc_average, tcF) (Yu et al. 2014; Xie et al. 2019; Amjad et al. 2022). Each category represents a distinct level of the combined impact of the first authors and corresponding authors in a research paper, reflecting the combined influence of these related observable variables.These related obesrvable variables encompasses the number of papers published (pNum_Max, pNum_average, pNumF) (Stremersch et al. 2015), the number of citations received by published papers (tc_Max, tc_average, tcF) (Yu et al. 2014; Xie et al. 2019; Amjad et al. 2022), the h-index (h_max, h_average, HIF) (Wang et al. 2012; Wang et al. 2019; Xie et al. 2019), centrality measures in the collaboration network (auCDF, degree_max, degree_average), eigenvector centrality (Eigenvector_centrality_Max, Eigenvector_centrality_F, Eigenvector_centrality_average) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021), and the number of authors per paper (authors) (Yu et al. 2014; Bornmann and Leydesdorff 2015; Xie et al. 2019).

inst_factors represents a latent variable that integrates multiple observable variables related to the institutions influence of the paper authors. *inst_factors* is categorized into four categories based on a comprehensive assessment of various observable variables, including centrality measures of research institutes in the collaboration network (inst_degree_average, inst_degree_max, inst_degree_F) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021), eigenvector centrality of research institutes (inst_Eigenvector_centrality_average, inst_Eigenvector_centrality_max, inst_Eigenvector_centrality_F) (Zhang et al. 2021), and the number of research institutes (institution) (Wang et al. 2019; Zhang et al. 2021). Each category represents a distinct level of the combined impact of the institutions affiliated with the authors in a research paper, reflecting the combined influence of these related observable variables. Table 1 presents the latent variables and their corresponding observable variables.

Finally, research work affects the paper quality, we utilize Normalized Citation Impact (CNCI) to evaluate the quality of academic papers. This is because, as Li

(2019) pointed out, current academic paper evaluation methods mainly characterize from the perspectives of impact and innovativeness. According to Tahamtan et al. (2016), creativity and novelty are features influencing internal factors of papers, and we classify paper innovativeness into *paper_factors*. Based on the extensive use of CNCI by scholars in the field of scientometrics to measure paper impact, and the fact that conclusions based on this metric are generally considered representative (Lei and Yan 2016; Ante 2022; Bornmann and Leydesdorff 2015), this paper only employs CNCI to evaluate the quality of academic papers.

Table 1. Latents Variable and their Corresponding Observable Variables in a BN.

Latent variables	*corresponding observable variables*	*meaning of corresponding observable variables*
paper_ factors	pnov	Paper novelty (Bu et al. 2021)
	pDisrupt	Paper disruption (Wu et al. 2019)
	refsNum	Number of references (Rigby 2013; Onodera and Yoshikane 2015; Xie et al. 2019)
	abER	Summary text readability (Stremersch et al. 2015; Lei and Yan 2016; Ante 2022)
	abatract_length	Summary text length (Vamclay 2013; Xie et al. 2019; Ruan et al. 2020)
	title_length	Title text length (Stremersch et al. 2015; Xie et al. 2019)
	key_words_length	keyword text length (Xie et al. 2019)
	refcitation_average	Average number of citations for references (Bornmann and Leydesdorff 2015; Xie et al. 2019)
	refcitation_sum	Total number of citations of references (Xie et al. 2019)
	Rank	CCF Rank (Qian et al. 2017)
aus_factors	pNum_Max	Number of published papers (max) (Stremersch et al. 2015)
	pNum_average	Number of published papers (average) (Stremersch et al. 2015)
	pNumF	Number of published papers (first author) (Stremersch et al. 2015)
	tcF	Total citations (first author) (Yu et al. 2014; Xie et al. 2019; Amjad et al. 2022)
	tc_Max	Total citations (max) (Xie et al. 2019; Amjad et al. 2022)
	tc_average	Average citations (Xie et al. 2019; Amjad et al. 2022)
	HIF	h-index (first author) (Wang et al. 2012; Wang et al. 2019; Xie et al. 2019)
	h_max	h-index (max) (Hurley et al. 2013; Xie et al. 2019)
	h_average	h-index (average) (Xie et al. 2019)
	auCDF	Co-authorship network centrality degree (first author) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	degree_max	Co-authorship network centrality degree (max) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	degree_average	Co-authorship network centrality degree (average) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	Eigenvector_centrality_Max	Co-authorship network eigenvector centrality (max)
		Co-authorship network eigenvector centrality (first author)
	Eigenvector_ce	Co-authorship network eigenvector centrality (average)

	ntrality_F	(Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	Eigenvector_centrality_average	
	authors	Number of authors (Yu et al. 2014; Bornmann and Leydesdorff 2015; Xie et al. 2019)
inst_factors	institution	Number of institutes (Wang et al. 2019; Zhang et al. 2021)
	inst_degree_average	Cooperation network centrality degree (institute with average value)
	inst_degree_max	Cooperation network centrality degree (institute with maximum value)
	inst_degree_F	Cooperation network centrality degree (institute with first author value) (Didegah and Thelwall 2013; Xie et al. 2019; Zhang et al. 2021)
	inst_Eigenvector_centrality_average	Cooperation network eigenvector centrality (institute with average value)
	inst_Eigenvector_centrality_max	Cooperation network eigenvector centrality (institute with maximum value)
	inst_Eigenvector_centrality_F	Cooperation network eigenvector centrality (institute with first author value) (Zhang et al. 2021)

Structural Constraints

The potential structures of the BN are illustrated in Figure 3, consisting of two parts: local structure and global structure. The local structure is the latent class model mentioned earlier. The structure constraints for global network include: (1) Authors and institutions may be able to reference each other; (2) Authors and institutions can reference paper features, but not vice versa; (3) Authors, institutions, and paper features may be able to reference CNCI, but not vice versa. Given these constraints, there are a total of 58 compliant potential network structures. Our goal is to use a BN learning algorithm to select the structure that best matches the data distribution and estimate the parameters accordingly. Fig. 3 demonstrates a potential network structure. The *aus_factors* directly influences (points to) the *paper_factors*, *inst_factors*, and *CNCI*. Similarly, the *inst_factor*, in turn, directly influences (points to) the *aus_factors*, *paper_factors*, and *CNCI*. Moreover, the *paper_factors* directly influences (points to) *CNCI*.

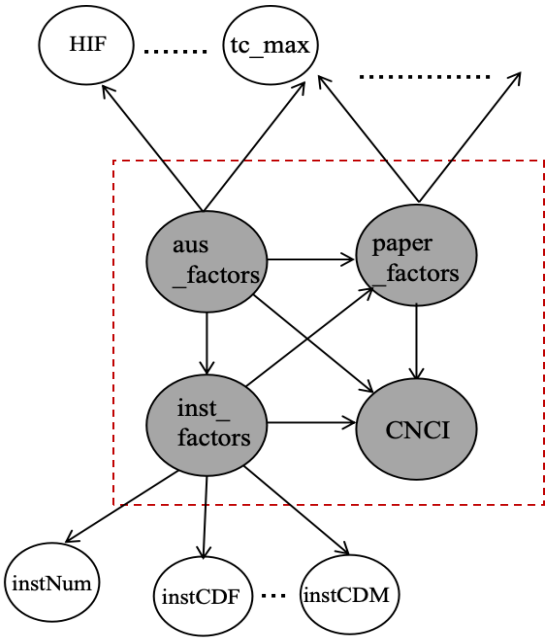


Figure 3. A potential network structure.

Bayesian Network (BN) with latent variables construction

Data preparation

The paper utilizes the Aminer paper dataset (Tang, 2008) as the foundational data, which is employed for computing the observable variables in Table 1. The Aminer dataset comprises a comprehensive collection of academic research papers and citation relationships, and it has been extensively utilized in various research endeavors related to academic research evaluation (Abramo et al., 2019; Amjad et al., 2022; Shao et al., 2022; Song et al., 2018). It has been employed in numerous studies associated with academic research evaluation. The dataset provides detailed information including paper identification number (*id*), title (*title*), publication date (*year*), author details (including identification numbers (*_id*), names (*name*), institutional affiliations (*org*), and institutional identification numbers (*gid*)), publication venues (including publication identification numbers (*_id*), publication names (*raw*)), keywords, abstract, citation counts (*n_itation*), reference number, and complete citation relationships among papers. Based on this information, we can compute the values of all the listed variables in Table 1 except for CCF Rank (https://www.ccf.org.cn/c/2019-12-01/666146.shtml). Regarding CCF Rank, given that our dataset covers academic journals and conferences in the field of computer science, and considering the significant influence of CCF rankings along with the absence of metrics such as JIF for conference papers, we introduce CCF Rank as a substitute for JIF Rank. This approach aids in accurately reflecting the importance of the papers.

In addition to CCF Rank, prior to BN learning, the factor values should be discretized into states. The values of CCF Rank can be A, B, or C. The discretization rule for other factors utilizes the equal-width binning method, whereby variables are sorted

in ascending order according to numerical values and divided into four equal intervals.

As shown in Table 2, we give the reasons for the missing values of various factors in the Aminer data. For conducting latent variable class learning based on K-Means, all attributes (i.e., the variables in Table 1) must have values, which requires each record to be complete. Hence, 96,760 complete records are used as the data source for the BN learning.

Table 2. Reasons for missing Factors values.

Factors	Missing reasons
auCDF, Eigenvector_centrality_F, HIF, tcF, pNumF	The first author identification number (First_aus_id) is missing
authors,degree_max,degree_average, Eigenvector_centrality_Max Eigenvector_centrality_averge h_max, h_average, tc_Max, tc_average pNum_Max, pNum_average	The entire author field is missing
inst_degree_F, inst_Eigenvector_centrality_F	The first author's institution identification number is missing
institution, inst_degree_average, inst_degree_max inst_Eigenvector_centrality_average inst_Eigenvector_centrality_max	Author's institution field is missing
abER, abatract_length	Summary field missing
title_length, key_words_length	Reference field missing
refsNum, refcitation_average, refcitation_sum, pnov, pDisrupt	(1) Reference field is missing. (2) Lack of real reference relationships
CNCI	The number of citations field is missing
Rank	Lack of publications. Only the grades (A, B, C) of journals and conferences in the CCF catalog are retained in publications.

Learning algorithm

Based on the given data, we employ the BN learning algorithm to learn its optimal structure and parameters. The input data consists of the observable variables D = [*aus*, *inst*, *paper*], where *aus* represents the list of observable variables corresponding to the latent variable *aus_factors*, *inst* represents the list of observable variables corresponding to the latent variable *inst_factors*, and *paper* represents the list of observable variables corresponding to the latent variable *paper_factors*. The cardinality of the latent variables is determined to be 4 based on expert knowledge (drawing on journal classification). The output includes the complete dataset as well as the optimal structure and its parameters. First, the algorithm uses the K-means algorithm to cluster the observable variables, obtaining categories corresponding to the latent variables (line 1-4). Next, all possible candidate structures are generated based on the structure constraints (line 5). Then, a scoring function is used to evaluate these candidate structures to find the optimal structure (line 6-11). Finally, the corresponding parameters for the optimal candidate structure are calculated (line 12-13). We implemented Algorithm 1 using the sklearn package(https://scikit-learn.org) and the pgmpy package(https://pgmpy.org/). The sklearn package covers almost all mainstream machine learning algorithms. It provides wrappers for common machine learning algorithms, including classification, regression, clustering, and dimensionality reduction. pgmpy is a pure python implementation for the BN with a focus on structure learning, parameter estimation, approximate and exact inference.The data preparation procedures (data preprocessing, variable value

calculation and discretization) were also implemented in Python. Since the K-means clustering algorithm requires a complete dataset to learn the latent variable categories, some data values may be unavailable due to missing data. Therefore, we removed the data with missing values and used the complete dataset to learn the latent variable categories. The Bayesian Information Criterion (BIC) (Schwarz, 1978) is used as the scoring metric to evaluate whether a candidate model is suitable for a given dataset. According to the structural constraints given above, we obtained a total of 58 candidate structure sets. As shown in Figure 4, we present some candidate structure sets. The structure with the highest score is considered the optimal structure. Once the optimal structure is determined, the network parameters can be easily learned from the data using the Maximum Likelihood Estimation (Zhang & Guo, 2006).

In the end, we obtained two optimal structures as shown in Figure 5. This model suggests that the reputation of institutions (*inst_factors*), the capability/influence of authors *(aus_factors*), and the features of the papers (*paper_factors*) all have impact on *CNCI*, and there is no single factor that can isolate the influence of another factor on the *CNCI*. Further, the results indicate that there is no explicit directional relationship between the influence of authors and institutions, meaning it is not clear whether the influence of authors determines the influence of institutions, or vice versa. Furthermore, the two optimal models are Markov equivalent (Zhang & Guo, 2006), which means they share the same probabilistic implications. Therefore, in subsequent analyses, as shown in Figure 5, optimal model (a) will be employed for inference and analysis. The learned BN with latent variables is shown in Figure 6.

Algorithm 1: Structure and Parameter Learning Algorithm

input : the set of observed variables data samples D,the cardinality of latent variables N

output: The complete dataset $DATA$, the optimal structure $G = (V, E, P)$ `// P is the conditional Probability table`

$DATA \leftarrow []$;
for *each the observed variables data set corresponding to the latent variable* X *in* D **do**
  $X \leftarrow X$ based on cluster centers N and input data X using k-means;
  $DATA \leftarrow$ Append the updated X to the $DATA$
$G_c \leftarrow$ generate the candidate structure set according to constraints ; $E \leftarrow \emptyset$, $G \leftarrow (V, E)$;
$S_m \leftarrow$ calculate the structure score of G based on $DATA$;
for *each* G_i *in* G_c **do**
  $S_i \leftarrow$ calculate the structure score of G_i based on $DATA$;
  if $S_i > S_m$ **then**
    $S_m \leftarrow S_i$, $G \leftarrow G_i$;
$P \leftarrow$ evaluate P based on $DATA$ using Maximum Likelihood Estimation;
return $G = (V, E, P)$

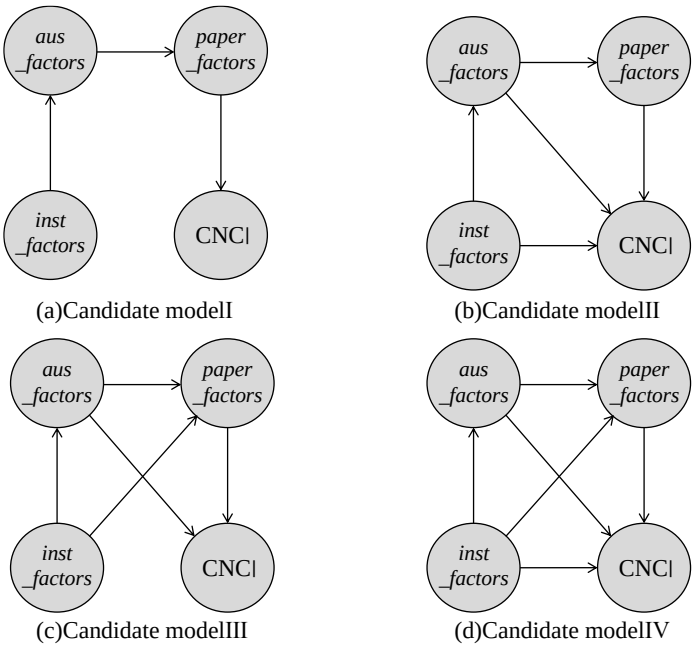


Figure 4. Some candidate models.

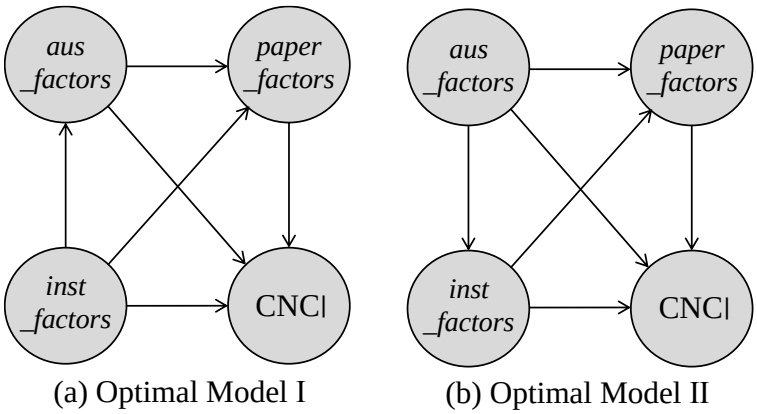


Figure 5. Optimal models.

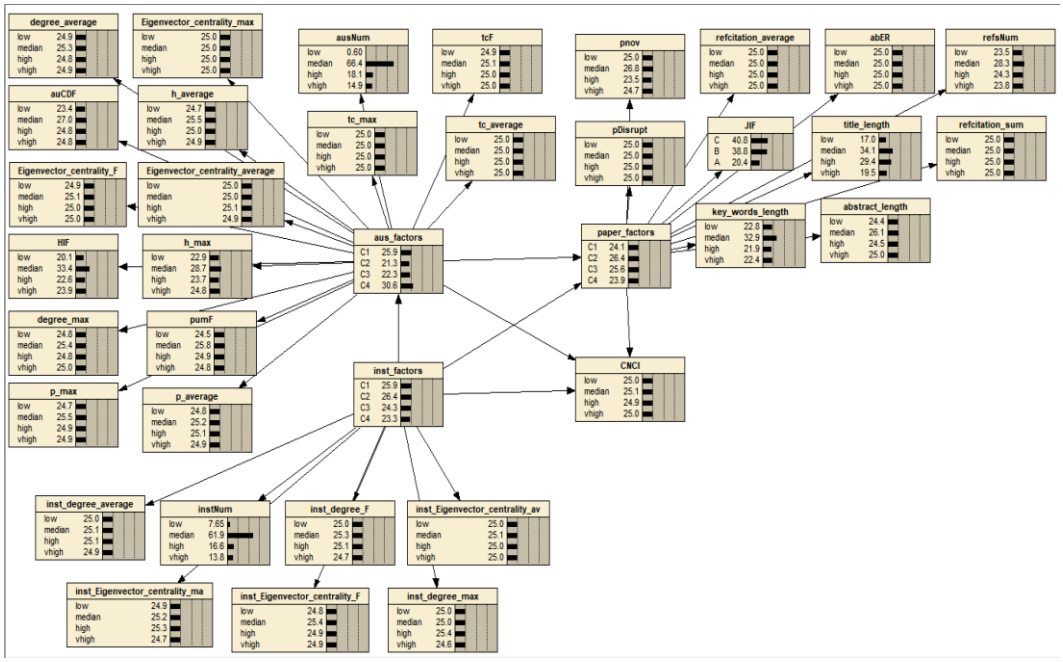
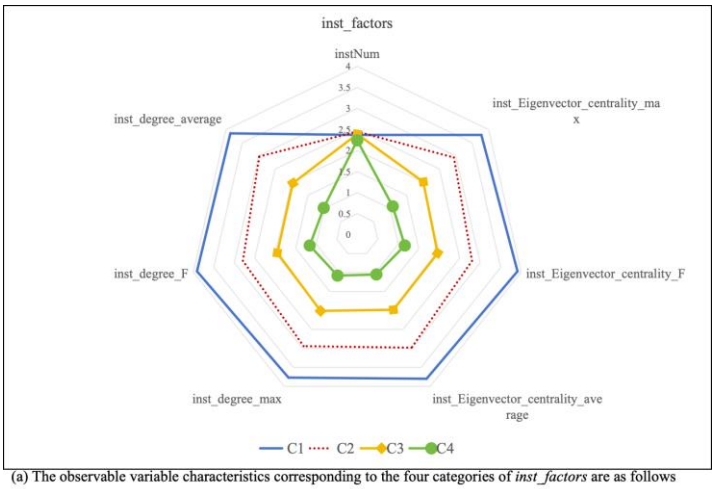


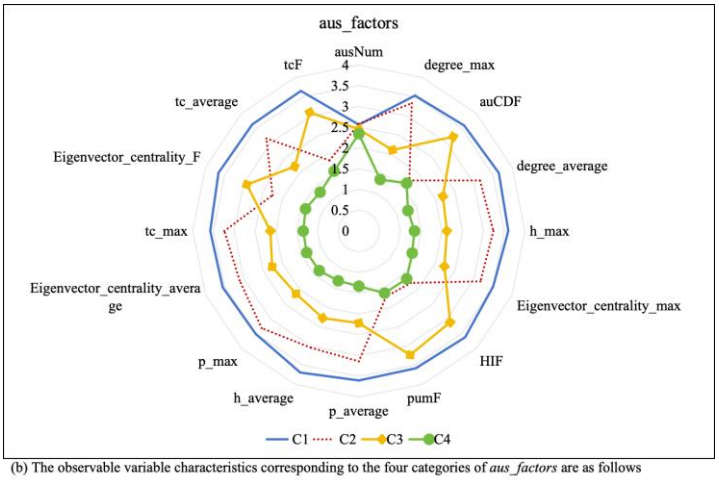
Figure 6. The learned BN with Latent Variables of Model (a).

Results

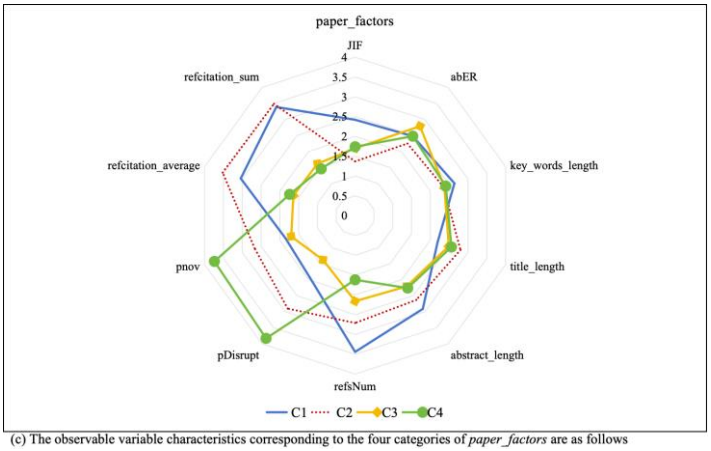
Based on the four categories (C1, C2, C3, C4) of the three latent variables (*inst_factors*, *aus_factors*, *paper_factors*), we calculated the mean values of the corresponding observable variables, as shown in Figure 7. Intuitively, one might expect a certain partial order relationship among the average values of the categories of the observable variables, allowing us, for example, to determine when comparing two categories of authors that one category is superior to another (in a certain sense). However, as illustrated in Figure 7(c), there is an overlap between categories C2 and C3 in terms of the average values of the observable variables corresponding to the four categories of *aus_factors*. This overlap arises from the fact that the data used to learn the BN is paper oriented. Papers are often authored by a group of scholars with varying characteristics (such as h-index), and authors from different clusters may exhibit certain intersections in terms of variable values. For instance, in a highly influential paper authored by three scholars, the h-index of each author might appear as (high, high, high), (high, high, low), or (high, medium, low), among others. That is, high-impact papers are not necessarily co-authored solely by high-impact authors, and similarly, low-impact papers may not be co-authored solely by low-impact authors. This scenario leads to the overlap between categories C2 and C3. The same situation also occurs in the *paper_factors* in Fig. 5(b). The latent variables in this paper are used to describe the overall influences of the categories of the corresponding observable variables as a whole, where these categories are learned from the combinations of the real situations implied in the real bibliometrics data.



(a) The observable variable characteristics corresponding to the four categories of *inst_factors* are as follows



(b) The observable variable characteristics corresponding to the four categories of *aus_factors* are as follows

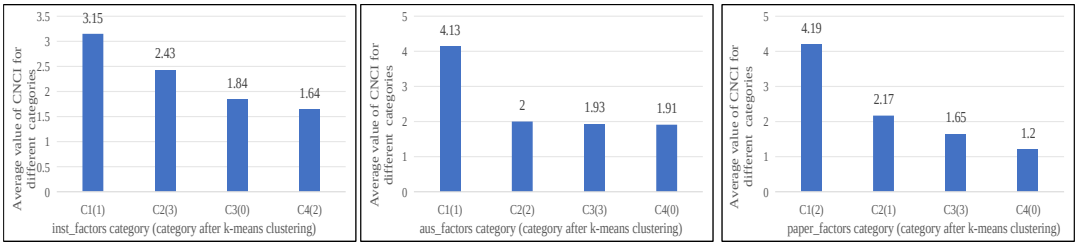


(c) The observable variable characteristics corresponding to the four categories of *paper_factors* are as follows

Figure 7. Differential characteristics presented by the four categories of latent variables in terms of observable variables.

To gain an understanding of the influence of *inst_factors*/*aus_factors*/*paper_factors* in each category, we calculated the average CNCI for papers corresponding to each category, as shown in Figure 8. Taking *inst_factors* as an example, as shown in

Figure 8(a), categories 1, 3, 0, 2 correspond to average *CNCI* values of 3.15, 2.43, 1.84, 1.64, respectively. To indicate varying influence among the categories and for ease of subsequent analysis, we named each category based on the ranking of their average CNCI values, with lower-numbered categories representing higher influence. Therefore, for *inst_factors*, categories 1, 3, 0, and 2 are named C1, C2, C3, and C4 respectively. Figure 8(b) and (c) show the situations of *aus_factors* and *paper_factors*.



(a) Average value of CNCI different inst_factors categories (b) Average value of CNCI different aus_factors categories (c) Average value of CNCI different paper_factors categories

Figure 8. Average CNCI values corresponding to the four categories of latent variables.

The necessity of adding latent variables

Using the approach taken by Sun et al. (2023), which sets different state combinations of each observed variable, for example, in the case of observable variables related to the authors, we can represent the situation of the authors in the paper and observe how these combinations affect the CNCI of the paper. However, although the combination of authors in some papers is different, the academic impact of the papers they publish is very similar. Setting observed variables can represent that a certain type of author combination can publish papers with high or low academic impact, but it is difficult to simultaneously represent the effect of several types of author combinations in producing similar academic impact (such as higher or lower academic impact). For example, when we set pNumF=*low* and h_max=*low*, the probability of CNCI from *low* to *vhigh* is 36.8%, 27.7%, 21.1%, 14.1%; when we set pNumF=*median* and h_max=*low*, the probability of CNCI from *low* to *vhigh* are 36.3%, 27.6%, 21.3%, and 14.7%. This shows that the academic influence of papers published by these two types of author combinations is similar, but it is impossible to express these two types of author combinations at the same time by setting observed variables. Furthermore, when the number of observed variables involved in author combinations is large, The situation will become more complicated, (such as setting auCDF, Eigenvector_centrality_F, HIF, tcF, pNumF, authors,degree_max,degree_average, Eigenvector_centrality_Max,Eigenvector_centrality_averge, h_max, h_average, tc_Max, tc_average, pNum_Max, pNum_average at the same time). Furthermore, setting different state combinations of each observed variable (Sun et al., 2023) fails to capture the combinations of author characteristics (impact) in actual, paper-oriented scenarios. As noted in the first paragraph of the Results section, we manually set author's h-index to the (*high*, *high*, *high*) state, representing the

expected combination of author characteristics that would produce papers with higher average impact (as indicated by the papers' average CNCI value). This assumption stems from the intuitive belief that a combination of (*high, high, high*) h-index values among authors is more likely to result in a paper with higher average impact. However, this method still fails to identify the combinations of author characteristics that contribute to producing papers with a higher average impact(as indicated by the papers' average CNCI value).

Using our approach,which sets a single latent variable, we can simplify the complex combination of observed variables and classify different combinations of authors that produce similar academic impact into the same category. As shown in Figure 8(b), taking *aus_factors* as an example, we divide the latent variables into 4 levels (C1, C2, C3, C4). Compared with the *aus_factors* of the C2, C3, and C4 categories, the *aus_factors* of the C1 category include various author combinations. What these author combinations have in common is that their published papers have the highest average academic impact. Moreover, the author characteristic combinations represented by each category of *aus_factors* is paper-oriented and reflects real scenarios. This fundamentally differs from the method in Sun et al. (2023), which represents the expected combinations of author characteristics. As stated in the first paragraph of the Result section, setting *aus_factors*=C1 captures combinations such as (*high, high, high*), (*high, high, low*), or (*high, medium, low*) in terms of the authors' h-indices. This reflects the actual author combinations in real papers and is paper-oriented. The latent variable helps clarify the author characteristic combinations that contribute to papers with higher average impact (as indicated by the papers' average CNCI value). This approach is fundamentally different from the method in Sun et al. (2023), which involves manually setting each author's h-index to (*high, high, high*) to represent the expected combination of author characteristics for producing papers with higher average impact (as indicated by the papers' average CNCI value).

Therefore, there is a distinction between the meanings of observable and latent variables. For example, in the case of authors, observable variables refer to authors with different levels of influence, such as those measured by h-index or the number of published papers, whereas latent variable(*aus_factors*) represents combinations of author characteristics corresponding to different papers average impact levels(as measured by the papers' average CNCI values). As shown in Figure 7(c), within the four categories of *aus_factors*, there is overlap between C2 and C3 in terms of the average values of observable variables. This non-hierarchical overlap, observed from the data perspective, suggests that author characteristic combinations corresponding to different paper average impact levels (as measured by the papers' average CNCI values) exhibit differences when compared to authors with varying levels of influence (ranging from *vhigh* to *low*).

Due to the differences in the meanings of observable and latent variables, it is clear that studying the interactions between observable variables differs significantly from studying the interactions between latent variables. For instance, in research involving institutions and authors, by jointly setting different states of observable variables (e.g., HIF = *high*, pNumF = *high*) and observing the distribution of institutions(e.g.,instCDM), interactions between observable variables typically focus

on how high-impact authors are distributed across institutions with different reputations. In contrast, by individually setting different states of latent variables (e.g., *aus_factors*=C1) and observing the distribution of *inst_factors*, interactions between latent variables tend to adopt a paper-oriented perspective, focusing on how author characteristic combinations in papers with higher average impact (as measured by the papers' average CNCI value) are distributed across institutional characteristic combinations with varying levels of paper average impact (as measured by the papers' average CNCI value).

Furthermore, latent and observable variables differ in how they contribute to understanding the interaction between paper impact and the factors that influence paper impact. For instance, by jointly setting different states of observable variables (e.g., HIF=*high*, pNumF=*high*) and observing the distribution of paper impact, this approach focuses on the paper impact distribution of papers written by high-impact authors. By individually setting different states of latent variables (e.g., *aus_factors*=C1) and observing the distribution of paper impact, this method focuses on the influence distribution of papers written by author characteristics combinations with higher average paper impact (measured by the average CNCI value of the papers). This distinction reflects the differing research perspectives and methodologies of observable and latent variables in paper impact analysis. In general, the interactions between latent variables and their relationship to paper impact differ significantly from the role of interactions between observable variables in influencing paper impact. The following section, "***Inferring the BN with Latent Variables***," will provide an example from the data perspective, analyzing the differences between observable and latent variables and exploring how interactions between latent variables affect paper impact.

Finally, by introducing latent variables, this method, compared to Sun et al. (2023), enables the study of interactions between latent and observable variables. By analyzing these interactions, it is possible to reveal the characteristic combinations of authors at different levels of paper average impact across the observable variable dimensions. The following section, ***Characteristics of Different Categories of the Latent Variables***, provides a more detailed analysis.

In summary, the introduction of latent variables not only simplifies complex combinations of observed variables, helping to classify author/institution/paper itself combinations with similar academic impact into the same category, but also represents a real, paper-oriented combination of multiple author/institution/paper characteristics. Compared to the method of Sun et al. (2023), this approach better captures the interactions between latent variables in real, paper-oriented scenarios, as well as the interaction between these latent variables and paper impact. Furthermore, the introduction of latent variables allows for the study of the interactions between latent and observed variables, revealing the characteristics of latent variables at different levels of paper average impact across various observed variable dimensions. This expands our understanding of Analysis of relationships between paper citations and their category influencing factors, which enhances the depth of the research in higher-level macroscopic perspectives.

Characteristics of different categories of the latent variables

Now let's take a detailed look at characteristics of different categories of the latent variables, and explore the relationship between these characteristics and CNCI. First, from the perspective of *inst_factors* clustering, as shown in Figure 7(a), from C4 *inst_factors* to C1 *inst_factors*, the degree of collaboration within the institution (*inst_ degree_F*, *inst_ degree_Max*) and the importance of the institution in the network (*inst_Eigenvector_centrality_F*, *inst_Eigenvector_centrality_max*) increase. At the same time, the average degree of collaboration within the organization (*inst_degree_average*) and the average importance of the organization in the network (*inst_Eigenvector_centrality_average*) also increase. However, there is almost no significant difference in the number of institutions in the 4 categories of *inst_factors*. This suggests that academic work is more likely to be cited when all participating authors are from institutions with higher degrees of collaboration and greater importance within the collaborative network.

Then, from the perspective of clustering based on *aus_factors*, as shown in Figure 7(b), the mean value of each observable variable in C1 *aus_factors* is the highest. In contrast, C4 *aus_factors* has the lowest mean value for each observable variable. In C2 and C3 *aus_factors*, the corresponding author has a higher mean value for each observable variable in C2, while in C3, the first author has a higher mean value for each observable variable. This indicates that academic work is more likely to be cited when all co-authors in a paper exhibit high level of each observable variable.

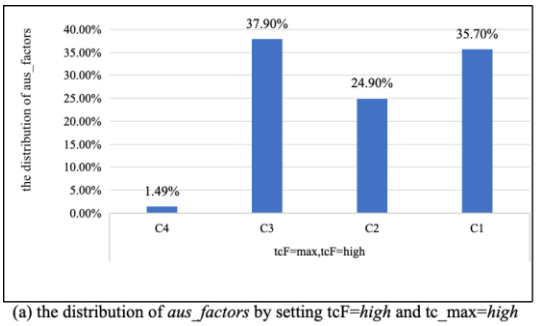
Further, from the perspective of clustering based on *paper_factors*, as shown in Figure 7(b), C1 *paper_factors* is generally published in the most influential publications. The number of references is the largest, and the average number of citations per reference and total number of citations of references are at a medium level. This shows that the research foundation of C1 *paper_factors* is relatively deep. In addition, C1 *paper_factors* tends to have the longest abstracts, the keywords with the largest number of words, and the most concise titles. However, C1 *paper_factors* is at a medium level of innovation and disruption. The number of references of C2 *paper_factors* is at a medium level, and the average number of citations per reference and total number of citations of references are the highest, indicating a deeper research foundation. Additionally, C2 *paper_factors* tends to have medium-length abstracts and the longest titles with the smallest number of keywords. It also exhibits moderate levels of innovation and disruption. Despite this, C2 *paper_factors* was published in the lowest impact journals. The number of references of C3 *paper_factors* is at a medium level, the average number of citations of references is at a medium level, the total number of citations is the lowest, and it lacks a deep research foundation. Additionally, C3 *paper_factors* tends to have the shortest abstracts, medium-length titles, and medium-level keywords. C3 *paper_factors* has the lowest level of innovation and disruption. C4 *paper_factors* is characterized by the highest level of disruption and innovation. This shows that high-impact works tend to be published in the highest-impact publications, with mid-range number of references, number of citations per reference, and total number of citations of references. They typically have the longest abstracts, the highest number of keywords, the most concise titles, and demonstrate a moderate level of innovation.

Finally, taking an overall perspective, hidden variables (*aus_factors*, *inst_factors*, *paper_factors*) possess their own characteristics, with some features having high impact while others have relatively lower impact. Through the analysis above, we conclude that in academic papers, (1) the higher the degree of collaboration of its institutional portfolio, the more important it is in the collaboration network and (2) the higher the influence of its author portfolio, the easier it is to be cited, in addition, (3) Additionally, the paper itself should be published in highly influential publications. It should have a moderate number of references, citations per reference, and total citations for references. Furthermore, it should feature a lengthy abstract, an extensive list of keywords, a succinct title, and display a moderate degree of innovation.

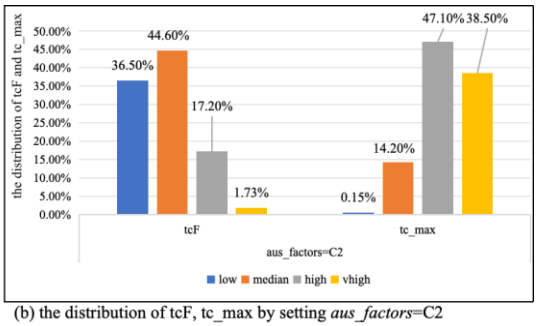
Inferring the BN with latent variables

This section will provide an example from the data perspective, analyzing the differences between observable and latent variables, observing the distributions of these three latent variables, and exploring how interactions between latent variables affect paper impact.

First, we will provide an example from the data perspective, analyzing the differences between observable and latent variables. Taking authors as a case study, we use observable variables tcF, tc_max to represent the number of citations of papers published by the first author and the corresponding author. We use the latent variable *aus_factors* to represent the combinations of author characteristics that result in different paper average impact levels (as measured by the papers' average CNCI values), as shown in Figure 9(a). When tcF=*high* and tc_max=*high*, the probability distribution of aus_factors from C4 to C1 is 1.49%, 37.90%, 24.90% and 35.70%. This indicates that authors with the same level of influence are not all categorized into the same group that produces papers with similar levels of average impact (as measured by the papers' average CNCI values). As shown in Figure 9(b), when aus_factors=C2, the probability distributions of tcF and tc_max from *low* to *vhigh* are 36.5%,44.60%, 17.20%,1.73% and 0.15%, 14.20%, 47.10%, 38.50% respectively. It can be observed that the probability distributions of tc_max and tcF from *low* to *vhigh* are not confined to a single state (i.e., the probability distribution is not 100% in one state). This indicates that the C2 category of *aus_factors* cannot be represented by a single joint setting of different states for tc_max and tcF. The setting of aus_factors=C2 is because the four levels of categories in aus_factors, namely C1, C2, C3, and C4, correspond to the four states of the observed variables: *vhigh*, *high*, *median*, and *low*. Through this example, it is clear that observable variables cannot represent latent variables through joint settings, and the meanings represented by latent variables are significantly different from those of observable variables.



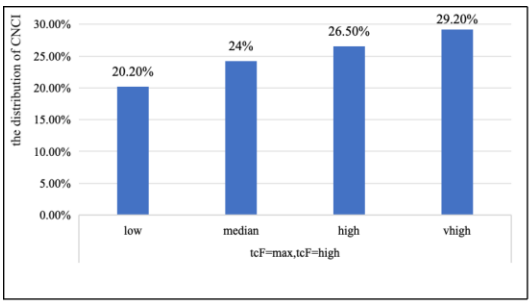
(a) the distribution of *aus_factors* by setting tcF=*high* and tc_max=*high*



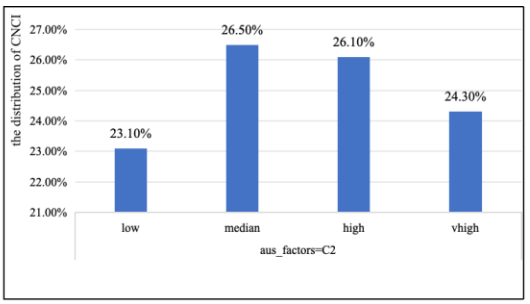
(b) the distribution of tcF, tc_max by setting *aus_factors*=C2

Figure 9. The distribution of aus_factors, tcF and tc_max.

It is evident that the interactions between observable variables and CNCI differ meaningfully from those between latent variables and CNCI. As shown in Figure 10(a), when setting tcF = high and tc_max = high, the probability distribution of CNCI is 20.20%, 24%, 26.50%, and 29.20%, which reflects the CNCI distribution for papers authored by researchers with a high level of influence. In contrast, as shown in Figure 10(b), when setting aus_factors = C2, the probability distribution of CNCI is 23.10%, 26.50%, 26.10%, and 24.30%, representing the CNCI distribution for papers authored by researcher combinations with papers of relatively high average impact. These two distributions have different meanings, and naturally, they result in different CNCI probability distributions, even for the same Aminer dataset.



(a) the distribution of CNCI by setting tcF=*high* and tc_max=*high*



(a) the distribution of CNCI by setting *aus_factors*=C2

Figure 10. The distribution of CNCI.

Then, we observe the distributions of these three hidden variables. As shown in Figure 11, in the Aminer paper dataset, the independent distribution of *C*1 *aus_factors* is 25.90%, the independent distribution of *C*1 *inst_factors* is 25.90%, and the independent distribution of *C*1 *paper_factors* is 24.1%. The independent distribution of *C*4 *aus_factors* is 30.6%, the independent distribution of *C*4 *inst_factors* is 23.30%, and the independent distribution of *C*4 *paper_factors* is 23.9%.

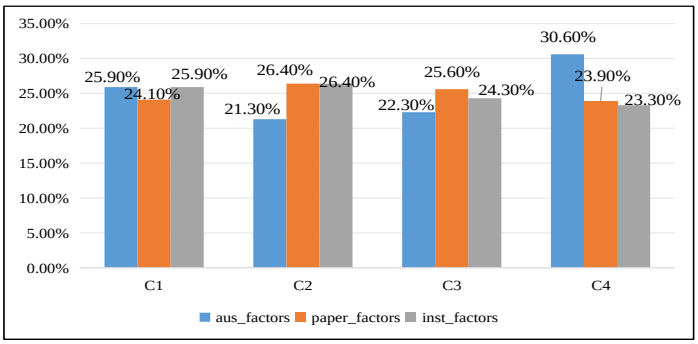
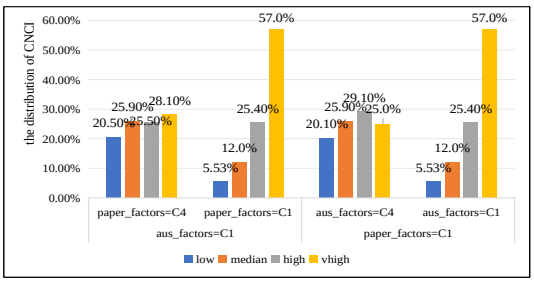
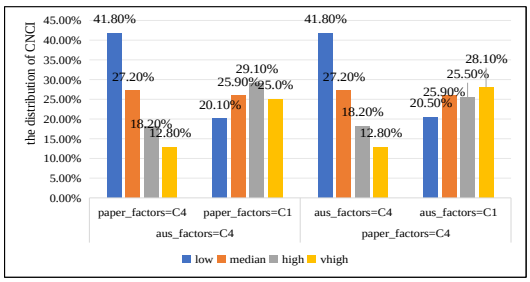


Figure 11. The independent distribution of latent variables.

We also can infer associations of latent variables with *CNCI* and the associations among themselves. Firstly, we first analyze which of *aus_factors* and *paper_factors* has a greater impact on *CNCI*. We find that the characteristics of authors on *CNCI* is slightly higher than the impact of internal features within the paper on *CNCI*. As shown in Figure 12(a), when we set *aus_factors*=*C*1 and change *paper_factors* from *C*4 to *C*1, the probability of *vhigh CNCI* increases from 28.10% to 57%, indicating a change of 28.9%. Similarly, when we set *paper_factors*=*C*1 and change *aus_factors* from *C*4 to *C*1, the probability of *vhigh CNCI* increases from 25% to 57%, indicating a change of 32%. This suggests that, compared to internal features within the paper, author characteristics has a slightly higher impact on the paper's influence (*CNCI*). As depicted in Figure12(b), when both *aus_factors* and *paper_factors* are set to *C*4 and the same operations are performed, the same conclusion is reached.



(a)the distribution of CNCI when setting aus_factors=C1 (paper_factors=C1) and changing paper_factors from C4 to C1 (aus_factors from C4 to C1)

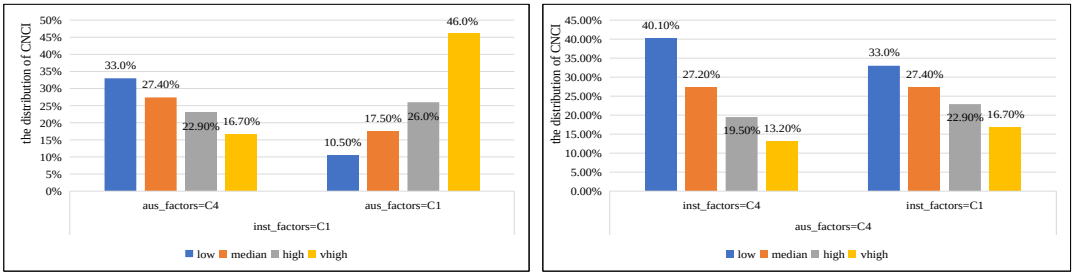


(b)the distribution of CNCI when setting aus_factors=C4 (paper_factors=C4) and changing paper_factors from C4 to C1 (aus_factors from C4 to C1)

Figure 12. Distribution of *CNCI* by setting various *aus_factors* and *paper_factors* values.

Next, we analyze the extent to which *aus_factors* and *inst_factors* affect *CNCI*. In addition, we also observed that, compared to institutional characteristics, author characteristics has a greater impact on the paper's influence. Furthermore, through inference, we speculate that within institutions, especially in *C*1 *inst_factors*, the most significant factor in altering the influence of a paper remains the prominence author characteristics within the institution. This underscores the idea that authors, rather than institutions, are fundamentally one of the most influential factors affecting the impact of a paper. In Figure 13(a), when *inst_factors* are fixed at *C*1 and *aus_factors* are changed from *C*4 to *C*1, there is a significant increase in *vhigh*

CNCI (16.7% → 46%). In Figure 13(b), when *aus_factors* are fixed at *C*4 and *inst_factors* are changed from *C*4 to *C*1, there is only a minor increase in *vhigh CNCI* (13.2% → 16.7%). This suggests that within *inst_factors*, especially in *C*1 *inst_factors*, *aus_factors* remains the primary factor in altering the impact of a paper.



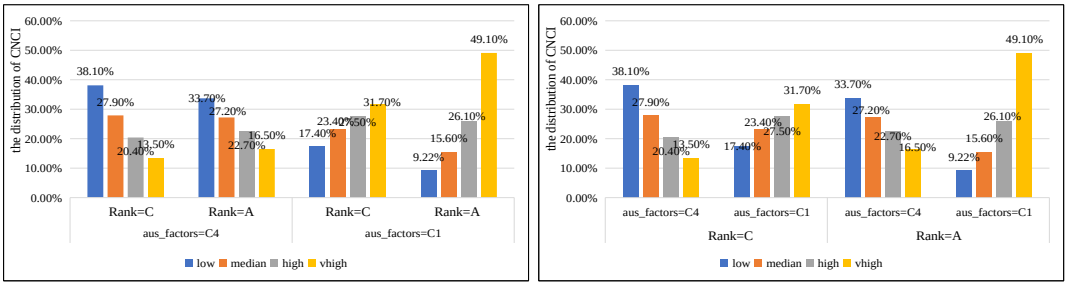
(a) the distribution of CNCI by setting inst_factors=C1 and changing aus_factors=C4 to C1

(b) the distribution of CNCI by setting inst_factors=C4 and changing aus_factors=C4 to C1

Figure 13. Distribution of CNCI by fixing inst_factors and set various aus_factors values.

In addition, many scholars have observed a significant positive correlation between the influence of the publications where papers are published and *CNCI* (Xie et al. 2019; Stegehuis et al., 2015; Didegah and Thelwall 2013). Therefore, we also analyzed the extent to which *aus_factors* and *Rank* affect *CNCI*. We also found that, compared to the *Rank*, author characteristics has a greater impact on *CNCI*. In Figure 14(a), when *aus_factors* are set to *C*4 and *Rank* is changed from *C* to *A*, the probability of *vhigh CNCI* increases from 13.5% to 16.5%, with a relatively small increase. In Figure 14(b), when *Rank* is set to *C* and *aus_factors* are changed from *C4* to *C1*, the probability of *vhigh CNCI* increases from 13.5% to 31.7%, indicating a relatively large increase. This suggests that, compared to *Rank*, *aus_factors* have a greater influence on the impact of the paper.

In conclusion, author characteristics is the most critical factor influencing *CNCI*. Compared to the intrinsic features of papers, the influence of author characteristics on *CNCI* is slightly higher. Within institutions, especially in *C*1 *inst_factors*, the most significant determinant of *CNCI* remains the author characteristics within the institution. Furthermore, in comparison to the *Rank*, the influence of author characteristics on *CNCI* is more pronounced.



(a) the distribution of CNCI by setting aus_factors=*C4*(*C1*) and changing *Rank* from *C* to *A*

(b) the distribution of CNCI by setting *Rank*=*C*(*A*) and changing aus_factors from *C4* to *C1*

Figure 14. Distribution of CNCI when different combinations of aus_factors and inst_factors are set.

Conclusion

In this paper, we adopt a BN with latent variables as the knowledge framework, using latent variables to describe characteristics of different aspects (i.e., institution, author and paper aspects) as a whole, so that interactions among latent category factors as well as observable factors can be analyzed. We use the K-means algorithm to acquire categories of latent variables and use constraint-based scoring approach to learn the BN. We analyzed how the introduction of latent variables provides new perspectives compared to using only observable variables, and conducted corresponding analyses, resulting in certain conclusions.

Leveraging BN with latent variables for inference has allowed us to derive the similar conclusions presented in Sun et al., (2023). However, the inclusion of latent variables has yielded more insights. For instance, within certain institutions, even in *C*1 *inst_factors*, author characteristics remain the primary factor influencing the impact of a paper. Compared with conclusion that authors have greater influence than institutions in Sun et al., (2023), our findings provide a deeper understanding of the interaction between institutions, authors, and CNCI. Additionally, we have uncovered some novel insights, such as from the perspective of papers, author characteristics are the key factors influencing CNCI, surpassing institutional features and paper content.

The data used for the BN construction comes from the Aminer dataset, implying that the research results of this paper are generally applicable to the field of computer science. Exploring the different models or pathways in different fields would be worthwhile in the future. Although this paper comprehensively uses latent variables to represent institutional factors, author factors, and internal paper features, concepts such as institutional influence, scholarly achievements, and paper innovation are complex. We only utilize a portion of bibliometric indicators to represent them, which may result in an incomplete understanding of domain knowledge.

In this study, we ignore the impact of time factor on citations and their categories. However, the temporal factor is crucial to understanding the interactive relationships between paper citations and their category influencing factors . In subsequent research, we will need a framework to study the dynamic interactive relationships between paper citations and their category influencing factors.

Acknowledgments

The work is supported by the Literature and Information Capacity Building Project of Chinese Academy of Science (No. E3291106).

References

Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting publication long-term impact through a combination of early citations and journal impact factor. Journal of Informetrics, 13(1), 32-49.

Amjad T, Shahid N, Daud A, et al. Citation burst prediction in a bibliometric network[J]. Scientometrics, 2022, 127(5): 2773-2790.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. American sociological review, 829-846.

Bornmann, L. (2011). Scientific peer review. Annual review of information science and technology, 45(1), 197-245.
Bornmann, L., & Leydesdorff, L. (2015). Does quality and content matter for citedness? A comparison with para-textual factors and over time. Journal of Informetrics, 9(3), 419-429.
Boyd, B. K., Finkelstein, S., & Gove, S. (2005). How advanced is the strategy paradigm? The role of particularism and universalism in shaping research outcomes. Strategic Management Journal, 26(9), 841-854.
Bu, Y., Waltman, L., & Huang, Y. (2021). A multidimensional framework for characterizing the citation impact of scientific publications. Quantitative Science Studies, 2(1), 155-183.
Chickering, D. M. (2002). Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov), 507-554.
Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. J. Mach. Learn. Res., 15(1), 3741-3782.
Daly, R., Shen, Q., & Aitken, S. (2011). Learning Bayesian networks: approaches and issues. The knowledge engineering review, 26(2), 99-157.
Didegah, F., & Thelwall, M. (2013). Which factors help authors produce the highest impact research? Collaboration, journal and document properties. Journal of informetrics, 7(4), 861-873.
Elidan, G., & Friedman, N. (2013). Learning the dimensionality of hidden variables. arXiv preprint arXiv:1301.2269.
Elidan, G., Lotner, N., Friedman, N., & Koller, D. (2000). Discovering hidden variables: A structure-based approach. Advances in Neural Information Processing Systems, 13.
Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., & Sebag, M. (2018). Learning functional causal models with generative neural networks. Explainable and interpretable models in computer vision and machine learning, 39-80.
He, C., Yue, K., Wu, H., & Liu, W. (2014, November). Structure learning of bayesian network with latent variables by weight-induced refinement. In Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning (pp. 37-44).
Hurley, L. A., Ogier, A. L., & Torvik, V. I. (2013). Deconstructing the collaborative impact: Article and author characteristics that influence citation count. Proceedings of the American Society for Information Science and Technology, 50(1), 1-10.
Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited−article, author, or journal?. Academy of management journal, 50(3), 491-506.
Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What causes a management article to be cited—article, author, or journal?. Academy of management journal, 50(3), 491-506.
Kalisch, M., & Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. Journal of Machine Learning Research, 8(3).
Kan, Y., Yue, K., Wu, H., Fu, X., & Sun, Z. (2022). Online learning of parameters for modeling user preference based on bayesian network. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 30(02), 285-310.
Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.
Kong, H., & Wang, L. (2023). Flexible model weighting for one-dependence estimators based on point-wise independence analysis. Pattern Recognition, 139, 109473.

Kulis, B., & Jordan, M. I. (2011). Revisiting k-means: New algorithms via Bayesian nonparametrics. arXiv preprint arXiv:1111.0352.
Latour, B. (1987). Science in action: How to follow scientists and engineers through society. Harvard university press.
Leimu, R., & Koricheva, J. (2005). What determines the citation frequency of ecological papers?. Trends in ecology & evolution, 20(1), 28-32.
McCabe, M. J., & Snyder, C. M. (2014). Identifying the effect of open access on citations using a panel of science journals. Economic inquiry, 52(4), 1284-1300.
Mingers, J., & Xu, F. (2010). The drivers of citations in management science journals. European Journal of Operational Research, 205(2), 422-430.
Moed, H. F., & Garfield, E. (2004). In basic science the percentage of "authoritative" references decreases as bibliographies become shorter. Scientometrics, 60, 295-303.
Mourad, R., Sinoquet, C., Zhang, N. L., Liu, T., & Leray, P. (2013). A survey on latent tree models and applications. Journal of Artificial Intelligence Research, 47, 157-203.
Podsakoff, P. M., MacKenzie, S. B., Bachrach, D. G., & Podsakoff, N. P. (2005). The influence of management journals in the 1980s and 1990s. Strategic management journal, 26(5), 473-488.
Qi, Z., Yue, K., Duan, L., Hu, K., & Liang, Z. (2022). Dynamic embeddings for efficient parameter learning of Bayesian network with multiple latent variables. Information Sciences, 590, 198-216.
Ramsey, J., Glymour, M., Sanchez-Romero, R., & Glymour, C. (2017). A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. International journal of data science and analytics, 3, 121-129.
Rigby, J. (2013). Looking for the impact of peer review: does count of funding acknowledgements really predict research impact?. Scientometrics, 94(1), 57-73.
Rigby, J. (2013). Looking for the impact of peer review: does count of funding acknowledgements really predict research impact?. Scientometrics, 94(1), 57-73.
Ruan, X., Zhu, Y., Li, J., & Cheng, Y. (2020). Predicting the citation counts of individual papers via a BP neural network. Journal of Informetrics, 14(3), 101039.
Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.
Shao, Z., Zhao, R., Yuan, S., Ding, M., & Wang, Y. (2022). Tracing the evolution of AI in the past decade and forecasting the emerging trends. Expert Systems with Applications, 209, 118221.
Song, K., Yue, K., Wu, X., & Hao, J. (2021). An efficient approach for parameters learning of bayesian network with multiple latent variables using neural networks and p-em. In Collaborative Computing: Networking, Applications and Worksharing: 16th EAI International Conference, CollaborateCom 2020, Shanghai, China, October 16–18, 2020, Proceedings, Part I 16 (pp. 357-372). Springer International Publishing.
Song, Y., Situ, F., Zhu, H., & Lei, J. (2018). To be the Prince to wake up Sleeping Beauty: The rediscovery of the delayed recognition studies. Scientometrics, 117, 9-24.
Stegehuis, C., Litvak, N., & Waltman, L. (2015). Predicting the long-term citation impact of recent publications. Journal of informetrics, 9(3), 642-657.
Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. International Journal of Research in Marketing, 32(1), 64-77.
Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. International Journal of Research in Marketing, 32(1), 64-77.

Sun, M., Ma, T., Zhou, L., & Yue, M. (2023). Analysis of the relationships among paper citation and its influencing factors: a Bayesian network-based approach. Scientometrics, 128(5), 3017-3033.
Sun, M., Yue, M., & Ma, T. (2023). Differences between journal and conference in computer science: a bibliometric view based on Bayesian network. Journal of Data and Information Science, 8(3), 47-60.
Tahamtan, I., & Bornmann, L. (2018). Core elements in the process of citing publications: Conceptual overview of the literature. Journal of informetrics, 12(1), 203-216.
Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008, August). Arnetminer: extraction and mining of academic social networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 990-998).
Wang, F., Fan, Y., Zeng, A., & Di, Z. (2019). Can we predict ESI highly cited publications?. Scientometrics, 118, 109-125.
Wang, M., Wang, Z., & Chen, G. (2019). Which can better predict the future success of articles? Bibliometric indices or alternative metrics. Scientometrics, 119, 1575-1595.
Wang, Y., Zhang, N. L., & Chen, T. (2008). Latent tree models and approximate inference in Bayesian networks. Journal of Artificial Intelligence Research, 32, 879-900.
Wu, C. J. (1983). On the convergence properties of the EM algorithm. The Annals of statistics, 95-103.
Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. Nature, 566(7744), 378-382.
Wu Xinran & Yue Kun. (2023). Bayesian network learning method with latent variables: Research review. Journal of Yunnan University (Natural Science Edition) (02), 298-313.
Xie, J., Gong, K., Li, J., Ke, Q., Kang, H., & Cheng, Y. (2019). A probe into 66 factors which are possibly associated with the number of citations an article received. Scientometrics, 119, 1429-1454.
Yue, K., Wu, X., Duan, L., Qiao, S., & Wu, H. (2020). A parallel and constraint induced approach to modeling user preference from rating data. Knowledge-Based Systems, 204, 106206.
Zhang, L., & Guo, H. (2006). Introduction to Bayesian Networks. Science Press.
Zhu, S., Ng, I., & Chen, Z. (2019). Causal discovery with reinforcement learning. arXiv preprint arXiv:1906.04477.

https://doi.org/10.51408/issi2025_049

Are Citation Context Information Stronger Related to Peer Ratings Than Citation Counts? A Descriptive Analysis

Paul Donner[1], Stephan Stahlschmidt[2], Robin Haunschild[3], Lutz Bornmann[4]

[1]*donner@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW),
Schützenstrasse 6a, 10117 Berlin (Germany)

[2]*stahlschmidt@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW),
Schützenstrasse 6a, 10117 Berlin (Germany)
University of Granada, Unit of Computational Humanities and Social Sciences (U-CHASS), EC3 Research Group,
Campus Universitario de Cartuja, 18071 Granada (Spain)

[3]*R.Haunschild@fkf.mpg.de*
Max Planck Institute for Solid State Research, Information Service,
Heisenbergstrasse 1, 70569 Stuttgart (Germany)

[4]*bornmann@gv.mpg.de, L.Bornmann@fkf.mpg.de*
Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society,
Hofgartenstr. 8, 80539 Munich (Germany)
Max Planck Institute for Solid State Research, Information Service,
Heisenbergstrasse 1, 70569 Stuttgart (Germany)

Abstract

In this study, we investigated whether citation context information is able to increase the validity of citation impact analyses to measure research quality compared to simple citation counts. We analyzed the statistical relationships of information extracted from structured citation context data in the Web of Science (Clarivate) such as the placement of citations within specific sections of an article with post-publication peer review quality ratings from Faculty Opinions (H1 connect), used as an external validity criterion for research quality. The study is based on publications in medicine and life sciences. Our findings reveal that quantitative metrics derived from citation contexts, particularly in-text citation counts, exhibit stronger correlations with expert evaluations compared to traditional citation counts. Consequently, integrating citation context data appears to improve the legitimacy and reliability of citation analyses as tools for assessing research quality.

Introduction

Implicit in conventional citation analysis, which is mostly an analysis of the times cited information from citation databases, is the assumption that all citations have equal value. A paper is cited or not – depending on its utility and merit. More detailed inspection of citations in scientific documents shows, however, that there are great differences in how literature is processed by the authors in their papers. Some papers are cited *en bloc* within a long list of other cited papers to demonstrate that there is literature available on a certain topic (mostly in the introduction of a paper) while other papers are discussed in great depth. Evidently, the former first category of cited paper has had less impact on the citing paper than the second one. Whereas the

traditional citation analysis – the times cited analysis – focuses on references in the reference list of a document, "an *in-text citation* is a *mention* of a reference within the full text of a document. A reference can be mentioned one or more times in a document. Each mention is an in-text citation" (Boyack, van Eck, Colavizza, & Waltman, 2018). It is one goal of citation context analysis (CCA) to further develop traditional citation analysis and to provide more detailed insights into the use and impact of publications. Recently, Clarivate has started to systematically provide citation context information in the Web of Science (WoS) for many citing publications. This data has now attained sufficient coverage that an initial analysis has become feasible.

In this study, we explored the possibilities of using Clarivate's citation context information for more meaningful citation analyses. We investigated if CCA can improve the validity of measuring research impact (as one important dimension of research quality) by bibliometric means. The reason for a hypothesized improvement in construct validity is that CCA goes beyond reference list citation counting, quantitatively and qualitatively. The quantitative extension lies in the counting of repeated in-text citations and in the information of how many papers are cited to support a particular statement. We used the latter information to calculate a score on the level of cited papers. This score indicates the proportion of in-text citations in which the paper was cited as the single cited reference to support a statement, rather than one of several. The qualitative extension consists of taking the position of citations (e.g., in certain sections) and the text surrounding citations into account (e.g., is there a direct use of the cited paper's content, or does the cited paper serve as a background reference for a certain topic).

We hypothesized that citation context information aggregated at the level of cited publications contains additional information relevant for the assessment of publications' research quality as higher quality research is used differently in citation contexts than lower quality research. Higher quality research is expected to be utilized more often as significant citations, rather than perfunctory citations, because their influence on the citing paper's author is assumed to be greater. This could manifest in different ways, such as a higher probability to be cited in specific paper sections (Cano, 1989; Maricic, Spaventi, Pavicic, & Pifat-Mrzljak, 1998; Tang & Safer, 2008), a higher probability to be used for certain purposes (Tang & Safer, 2008), more frequent mentions in the citing paper (Zhu, Turney, Lemire, & Vellino, 2015), or more frequently being cited as the only reference in a citation context, rather than being one of a string or block of references cited together in one context (Beck, Sandbulte, Neupane, & Carroll, 2018). CCA may offer a significant improvement of the underlying basis of citation analysis by moving from a superficial reference list analysis to a more sophisticated and data-rich in-text citation analysis.

In this study, we posed the research question whether CCA improves the measurement of research impact as one aspect of research quality compared to the usual citation count analysis. To answer this question, we compared peer ratings of focal papers on the platform Faculty Opinions (FO, provided by H1, https://connect.h1.co), formerly F1000, with information derived from the citation

contexts of focal papers in citing papers indexed in the WoS. Since peer ratings may be the best way of assessing the quality of focal papers (Bornmann, 2011), the correlation of the ratings with simple citation counts on the one hand and outcomes of the CCA on the other hand may reveal possible improvements by the consideration of citation context information in enriched citation analyses compared to simple citation counting.

Datasets and methods

Faculty Opinions dataset

FO is a medicine and life sciences post-publication appraisal and recommendation service. FO expert members ('peers') rate papers on a 3-level ordinal scale ('good', 'very good', 'excellent') to express their perceived quality level of a paper. Note that neither low-quality nor ordinary quality publications are rated as such. Peers must regard contributions as good or better to recommend them for consideration in the FO database. They do so publicly under their own name within the FO subscription service and usually provide a concise explanation of the importance of rated publications. Given its unique nature as a large-scale dataset on concise peer reviews, FO data has been applied extensively in bibliometric and altmetric research and we refer to Williams (2017) for an in-depth description of the platform and resulting data (this description is still current although the operator changed).

H1 provided us with a dataset of 246,245 peer ratings of scientific publications from their service for this study, current as of November 2023. We excluded 282 records: FO members can provide a dissent rating which are exceedingly rare. These express disagreement with an existing recommendation but did not fit into the three-level quality scale and were therefore excluded. For each publication year from 2001 on, there are more than 3000 annual FO recommendations. The peak publication year was 2012 with over 16,000 recommendations. Papers received on average 1.2 recommendations and 16% of papers received more than one recommendation. The most common rating score was 'good', with 51%, while 39% of ratings were judged 'very good', and 10% were rated 'excellent'.

Using the official publication date of the rated paper and the date of its recommendation, we computed how long it typically took for a recommendation to be made. The average passed time is 222 days, with a standard deviation of 765 days. However, about 14% of the recommendations were posted before the recorded official publication date. Although the typically short time interval lets us assume that FO ratings are unlikely to be affected by citation count information searched by FO members, it is possible that citing authors were partially informed and influenced by FO ratings.

Web of Science citation context data

We use an April 2024 snapshot of WoS that includes the SCIE, SSCI, AHCI, CPCI-S, and CPCI-SSH and which is licensed through, and made available by, the German Kompetenznetzwerk Bibliometrie (Schmidt et al., 2024). Citation context data is available in the WoS since 2021 on a large scale under the feature name of Enriched

Cited References. This includes currently a numeric value between 0.0 and 1.0 for the relative position of the reference in the text of a paper, the original and a standardized section title, as well as the inferred reference function. Contrary to the section classification building upon the well-studied introduction, methods, results, and discussion (IMRaD) structure (Sollaci & Pereira, 2004), the citation functions constitute a classification developed by Clarivate with five classes: 'background', 'basis', 'differ', 'support', and 'discuss'.

Matching and resulting analytical dataset

We constructed an analytical dataset by matching WoS data with citation context information to FO data. As we wanted to study the associations of citation context variables with quality assessments, our study is necessarily limited to those publications for which any citation context data is available. This study therefore does not include any uncited document records. It also does not include citation information from citing publications without citation context data. We first restricted the dataset to publications of the years 2020, 2021, and 2022 as these currently have the best relative coverage of citation context data. The used citation context data were from citing publications of any publication years. We also limited the data to papers with the document type 'article', since papers with different document types can be cited differently (Lundberg, 2007). For this restricted WoS dataset, we continued with the matching to the FO data.

WoS and FO records were matched primarily by the DOI. For FO records without DOIs, matching was done by exact match on journal title, volume, issue, and first page. We also wanted to include additional papers that have not been recommended by FO members but have been published in the same journals as the recommended papers. For identifying the papers without FO rating, we selected all unrated WoS records of document type 'article' published between 2020 and 2022 in journals which ever had published a rated paper in the entire FO dataset. For the purposes of our study, the publications without any FO rating but published in these journals were assigned the rating level 'unrated'. Table 1 summarizes the numbers of records in the different datasets.

Table 1. Overview of datasets.

dataset	*records*
(1) WoS items ever cited with any citation context information	31,219,721 items
(2) WoS articles from 2020 to 2022 with citation context information	4,570,945 articles
(3) items with FO recommendations (publications with the same DOIs in WoS were discarded)	192,328 items, 246,245 recommendations
(4) matched data of (2) and (3)	13,617 articles, 15,771 recommendations
(5) analytical dataset: (4) extended with unrated publications	1,531,556 articles, 15,771 recommendations

Variables and statistics used

We processed the citation context data and calculated variables on the level of cited items as follows:

- Ordinary citation counts and number of in-text citations: For example, an item cited by three papers, which is referenced in these papers 2, 1, and 5 times, has a citation count of 3, but 8 in-text citations.
- Relative shares of citation contexts of normalized sections: We calculated for each cited item the relative shares of citation contexts of normalized sections, as defined by Clarivate ('introduction', 'methods', 'results', and 'discussion'). We additionally defined the section as 'missing' when no section information was available. For instance, a cited item with 5 in-text citations, of which 4 are in the introduction and 1 in the discussion section, would have variable values of 0.8 for share of introduction section, 0.2 for share of discussion section and 0.0 for the shares of the other categories.
- Relative shares of citation functions: In the same manner, the relative shares of citation functions, as defined by Clarivate ('discuss', 'background', 'basis', 'support', and 'differ') were calculated.
- Relative share of an item being cited as a single reference: A new variable was created for the relative share of an item being cited as a single reference: The share was calculated from the relative position data. References cited closely together within a citing paper were identified as those whose positions were within 1% of a paper's page of each other. This normalization for paper length in pages is necessary: A difference of, say, 0.05 on the 0.0 to 1.0 scale of a relative position is a very small distance for two references in a two-page paper but a large distance in a 40-page paper. The 1% of a page parameter value was found experimentally to provide satisfactory results by testing different parameter values on how well they identify multi-citation clusters in sample articles. The single reference share expresses what proportion of an item's citation contexts is not in such multi-reference citation contexts, usually a string of multiple references to support a single claim or statement. It quantifies the share of citation contexts in which an item is the only reference cited to support a statement.

For the descriptive analyses of associations between the citation context variables, the polyserial correlation coefficient was used, which is designed for the quantification of associations between ordered categorical and numeric variables.

Results

Table 2 shows the polyserial correlations between the citation context variables and FO ratings, in two variants. First, four ordinal levels ('unrated', 'good', 'very good', and 'excellent') were used. Second, we restricted the analyses to FO rated items (i.e., 'good', 'very good', and 'excellent'). By using the restriction to that subset, we can show more clearly which citation context variables could potentially differentiate

quality at the high end. Multiple ratings for one item were not aggregated but treated as independent observations, so this view on the dataset has more observations than publication records. The results in Table 2 reveal that the number of citations and in-text citations are moderately associated with better ratings taking into account cited publications without FO rating as a supplementary fourth quality level. When excluding unrated items, the number of in-text citations also exhibits slightly higher agreements with the FO ratings than the number of citations. The coefficients for the citation section, citation function, and share as single reference variables in the table are much smaller than those for the number of (in-text) citations. Size and direction of these coefficients are inconsistent across the two calculation variants, with the exception of the 'results' section and 'differ' function shares.

Table 2. Polyserial correlation coefficients between FO ratings and citation context variables.

		including unrated papers (n=1,533,710)	*excluding unrated papers (n=15,745)*
number of citations		0.44	0.07
number of in-text citations		0.47	0.08
share as single reference		0.00	0.01
citation section	introduction	−0.06	0.03
	results	0.06	0.05
	methods	−0.03	−0.01
	discussion	0.03	−0.07
	missing	0.02	0.01
citation function	discuss	0.05	−0.03
	background	−0.04	0.03
	basis	−0.02	0.01
	support	0.00	−0.02
	differ	−0.02	−0.04

In order to have a more detailed insight into the relationship of citation (context) variables and experts' ratings, average values of the citation context variables for the four quality rating levels are presented in Table 3. The average values show that only the relationships between rating categories and citations and in-text citations are monotonically increasing. Averages for section and function shares and shares as single references only differentiate in some cases when comparing unrated to rated levels, e.g., for the 'results' section or the 'discuss' function.

Table 3. Average values of (in-text) citations, citation context variables, and share as single reference across rating categories (n=1,533,710).

		FO rating level			
citation (context) variable		*unrated*	*good*	*very good*	*exceptional*
citations		6.2	26.2	29.4	76.8
number of in-text citations		10.0	43.2	50.0	125.2
share as single reference		0.46	0.46	0.47	0.46
share of citation section	introduction	0.43	0.37	0.37	0.39
	results	0.11	0.13	0.15	0.15
	methods	0.10	0.08	0.09	0.08
	discussion	0.34	0.39	0.37	0.35
	missing	0.02	0.03	0.03	0.03
share of citation function	discuss	0.38	0.43	0.42	0.41
	background	0.46	0.43	0.42	0.45
	basis	0.12	0.11	0.11	0.11
	support	0.04	0.04	0.04	0.04
	differ	0.00	0.00	0.00	0.00

Discussion

Using citation context data that is available in the WoS since 2021 on a large scale, we investigated in this study whether CCA enhances the validity of the measurement of research impact, a critical aspect of research quality, compared to traditional citation count analysis. We conducted a quantitative analysis comparing peer ratings of papers from the FO platform, with citation context information from the papers' citations indexed in the WoS. Given that peer ratings may be a superior measure of the papers' quality, examining the correlation between these ratings, and both simple citation counts and CCA outcomes, could highlight potential enhancements from integrating citation context information into citation analyses versus relying solely on citation counts.

Our investigations of the association of research quality, in terms of FO ratings, and variables derived from citation context information, have brought to light intriguing findings. In general, our results show that the number of in-text citations associate more strongly with FO ratings than regular citation counts. The number of in-text citations thus exhibit higher construct validity as a proxy variable for research quality than citation counts. On the other hand, the correlational analysis has not shown any clear associations of the other investigated citation context variables with FO ratings. This study is subject to some limitations. It is limited in scope to medicine and life sciences as covered by FO. The generalizability of our findings is difficult to assess due to well-known differences of citation practices across fields of science: "there are large field-level differences that are reflected in position within the text, citation interval (or reference age), and citation counts of references" (Boyack et al., 2018).

A technical limitation of our study is given by the limited availability of in-text citations. As in-text citations are much more frequent for recent citing years (at the time of this study), we focused on recent literature. It is not guaranteed that our results are transferable to older citing years.

Acknowledgments

Access to WoS bibliometric data has been supported via the German Kompetenznetzwerk Bibliometrie, funded by the Federal Ministry of Education and Research (grant number: 16WIK2101A). Stephan Stahlschmidt's contribution has been funded by the German Federal Ministry of Education and Research (grant number: 01PH20006B). We would like to thank H1 staff for providing data access and Clarivate for detailed descriptions of the citation context information.

Refereces

Beck, J., Sandbulte, J., Neupane, B., & Carroll, J. M. (2018). A study of citation motivations in HCI research. Retrieved December, 2, 2024, from https://osf.io/preprints/socarxiv/me8zd

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology, 45*, 199-245. doi: 10.1002/aris.2011.1440450112.

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics, 12*(1), 59-73. doi: 10.1016/j.joi.2017.11.005.

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science, 40*(4), 284-290. doi: 10.1002/(SICI)1097-4571(198907)40:4%3C284::AID-ASI10%3E3.0.CO;2-Z.

Lundberg, J. (2007). Lifting the crown: Citation z-score. *Journal of Informetrics, 1*(2), 145-154. doi: 10.1016/j.joi.2006.09.007.

Maricic, S., Spaventi, J., Pavicic, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the American Society for Information Science, 49*(6), 530-540. doi: 10.1002/(SICI)1097-4571(19980501)49:6%3C530::AID-ASI5%3E3.0.CO;2-8.

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., . . . Stahlschmidt, S. (2024). The Data Infrastructure of the German Kompetenznetzwerk Bibliometrie: An Enabling Intermediary between Raw Data and Analysis. Retrieved from https://doi.org/10.5281/zenodo.13935407 doi:10.5281/zenodo.13935407

Sollaci, L., & Pereira, M. (2004). The Introduction, methods, results, and discussion (IMRAD) structure: A fifty-year survey. *Journal of the Medical Library Association, 92*, 364-367.

Tang, R., & Safer, M. A. (2008). Author-rated importance of cited references in biology and psychology publications. *Journal of Documentation, 64*(2), 246-272. doi: 10.1108/00220410810858047.

Williams, A. E. (2017). F1000: An overview and evaluation. *Information and Learning Science, 118*(7/8), 364-371. doi: 10.1108/ILS-06-2017-0065.

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology, 66*(2), 408-427. doi: 10.1002/asi.23179.

https://doi.org/10.51408/issi2025_059

Assessment of a Research Funding Organization for International Mobility by Bibliometric Means. Implementation, Results and Challenges of Responsible Research Evaluation

Torger Möller[1], Philippe Dittmann[2]

[1]*moeller@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW),
Lange Laube 12, 30159 Hannover (Germany)

[2]*dittmann.extern@dzhw.eu*
German Centre for Higher Education Research and Science Studies (DZHW),
Schützenstraße 6a, 10117 Berlin (Germany)

Abstract

Using the German Academic Exchange Service (DAAD) as an example, the study examines which of the three bibliometric databases (Dimensions, Scopus and Web of Science) has the best coverage for funding acknowledgement information. Web of Science provides most comprehensive data on funding acknowledgements, followed by Scopus and Dimensions. A special feature of the DAAD is the promotion of global academic mobility. In this respect, it is the largest funder in the world. The publications funded by the DAAD are examined regarding their worldwide distribution, their degree of internationalization and their excellence rate. A logistic regression is applied to investigate which factors influence the excellence rate of DAAD-funded publications. For this purpose, the funding acknowledgement data is linked with data from the research funding organization. The results reveal that the excellence rate depends on the funded academic group (graduates, doctoral candidates, postdocs, and faculty members), the gender, and the country of origin and destination of the grantee. The paper concludes with a discussion of how the results should be treated in the context of responsible research evaluation.

Introduction

Research performing organizations, especially universities, have long been subject of evaluations that also use bibliometric data (e.g. ARWU "Shanghai" Ranking, QS World University Rankings, Times Higher Education (THE) World University Rankings). In addition, there are rankings that are mainly based on bibliometric data, e.g. the Leiden Ranking and the Scimago Institutions Ranking. Comparable rankings for research funders do not exist. There are various reasons for this.

For the investigation of research performing organizations, the publications are assigned to the organizations via the address affiliations. Address affiliations are an integral part of the scientific publication system and are thus included in bibliometric databases for a long time. In contrast, publications are assigned to research funding organizations via the acknowledgement section of the publications, in which not only research funders but also other institutions and individuals are acknowledged. The natural-language funding acknowledgement texts lack a uniform notation of both the name of the funding agency and the funding program. In their literature review on funding acknowledgements, Álvarez-Bornstein and Montesi call this a "lack of data

normalization" (Álvarez-Bornstein and Montesi 2021), which leads to misassignments of publications to respective funders.

To obtain a valid dataset for bibliometric analysis, extensive data cleaning of the natural language texts is necessary. Sirtes (2013) and Möller (2019), for example, found over six thousand name variants for the German Research Foundation (DFG) in just one publication year in the funding organization field of the Web of Science. Möller also shows that the number of spellings varies between different funding bodies. One reason for this is that research funders have a wide range of guidelines, from none to very detailed ones, on how grantees should acknowledge the source of funding. The "dirty" (Sirtes 2013) funding acknowledgement data requires a great effort in cleaning and normalization to conduct sophisticated bibliometric studies on research funders. Because of this effort, many studies focus on a single or a small number of research funders (e.g., Costas and Yegros-Yegros 2013; Meier et al. 2023; Möller, Schmidt, and Hornbostel 2016; Sirtes 2013; Wang, Jesiek, and Zhang 2024). In addition, there are a few studies that link funding acknowledgement data with data provided by the research funders. A first study focusing on the Austrian Science Fund (FWF) showed that only a portion of the publications listed in the final project reports provided to the agency could be identified in Web of Science via a funding acknowledgement analyses (Costas and Yegros-Yegros 2013; van Wijk and Costas-Comesaña 2012). It is well known that the Web of Science, as well as any other bibliometric database, does not cover all publications, but the fact that only 72% of publications from project final reports have a funding acknowledgement (Costas and Yegros-Yegros 2013) illustrates that funding acknowledgement analyses cannot identify all funded publications. However, it should be noted that the coverage of funding acknowledgement data has improved substantially in recent years (Clarivate Analytics 2022). The results of the previous FWF study published in 2012/13 are therefore somewhat outdated. A more recent study on the German Research Foundation (DFG) also uses data from final reports and compares these with funding acknowledgement information (Meier et al. 2023; Möller, Scheidt, and Meier 2024). 92% of publications mention the name of the DFG as the funding source. However, the grant number was only provided in 74% of cases (Möller et al. 2024).

This study builds on the above research strand by investigating funding acknowledgements and data provided by a funding agency. However, the aim is not to point out the differences between the two data sources as done previously. Instead, the bibliometric data on funding acknowledgement is supplemented by data from a research funding agency to carry out more sophisticated analysis. The object of the study is the German Academic Exchange Service (Deutscher Akademischer Austauschdienst, DAAD), which, according to its own statement, is the world's largest funding organization for the international exchange of students and academics (DAAD 2024a). Möller (2019) shows that most research funders only provide funding to academics from their home country. The originality of this study lies on the one hand, in linking bibliometric funding acknowledgement data with data from the research funder; on the other hand, in the international orientation of the funding body, which not only supports academics from Germany and their

international mobility, but also academics from a wide range of countries and their mobility.
After explaining the methodological approach, we investigate in the *Results* section the coverage of DAAD-funded publications in three bibliometric databases (Dimensions, Scopus and Web of Science). Then we focus on the internationality of the DAAD-funding and the impact they achieved. By linking bibliometric funding acknowledgement data with data from the research funder, we are capable to analyze to what extent the impact (excellence rate, PP top 10%) of the funded publications depends on the different DAAD-funding programs, the belonging to an academic group and the grantee's country of origin and destination. In the context of a responsible bibliometric evaluation, it is further explored and discussed to what extent bibliometric impact indicators do adequate justice the funding objectives of the different programs of the research funder.

Methods

Research funding organizations usually have extensive knowledge about their funded projects and grantees, including the amount of funding and the funding period. This knowledge forms the basis for the monitoring of funding and is published in annual reports and special evaluations (e.g., DAAD 2024b). However, there is often a lack of reliable information regarding the output of the funding, in particular which scientific publications are the outcome of research funding. One reason for this is that the funding recipients do not, or not completely, report the publications they have produced to the research funders. Publication notifications are made during the funding period or immediately afterwards in the final reports. Many publications appear years later and thus after the final report was submitted. The research funders are not informed about these publications. In addition, the lists of publications in final reports are usually unstructured, making an evaluation laborious. This effort is usually not feasible by the employees of research funding organizations.
In 2008, the bibliometric database Web of Science (WoS) began to include funding information for the first time. This was achieved by extracting the funding acknowledgement from the general acknowledgements section, in which colleagues, scientific institutions and research funding organizations are thanked for their support or financial assistance. Acknowledgements are short, unstructured texts in natural language written by the authors.
Many research funding organizations (e.g. the German Research Foundation (DFG), see Meier et al. 2023: 13ff) provide their funding recipients with detailed guidelines on how to indicate the funding source in a publication. However, the DAAD has no general standards in this regard, neither about the naming (Deutscher Akademischer Austauschdienst or German Academic Exchange Service) nor about the use of a grant number. The DAAD uses a personal code internally and in communication between the DAAD and its grantees. In only a few cases, this personal code was also mentioned in the funding acknowledgements.
Regardless of the specific requirements for how research funding should be indicated, acknowledging the funding source has become an established academic

publication practice. An online survey conducted by the German Centre for Higher Education and Science Research in 2016 showed that 94% of the scientists and scholars always or usually cited research funding (Möller et al. 2024: 1). Analyses of the funding context of publications are therefore a suitable instrument for examining the publication output of funding, even in the absence of specific guidelines of single funders.

In the context of this study, DAAD-funded publications are defined as publications that include a reference to the DAAD in the acknowledgements, e.g. "This study was funded by the German Academic Exchange Service (DAAD)". Text mining methods were used to identify DAAD-funded publications. 148 variations of the DAAD name were found. These include the two official German and English names (Deutscher Akademischer Austauschdienst and German Academic Exchange Service), the acronym DAAD and a wide range of grammatical forms of the official spellings. In addition, many "unofficial" spellings were found, for example, instead of German Academic Exchange Service, the terms German Academic Exchange Program/ Foundation/ Council or Office.

The result of the above search was quality-assured in a subsequent step. This involved checking whether the designations (especially from the unofficial spellings) really refer to the German Academic Exchange Service. Does a research funding organization with a similar name exists or does another research funding organization also use the abbreviation DAAD? The checks showed that the US Army Research Office has an extensive funding program that also uses the abbreviation DAAD. Almost one thousand publications that were initially identified only by the acronym DAAD were excluded from the final dataset during the quality assurance procedure. The findings of the analyses of this first data set are presented in the *Results* section. As a first step, the coverage of DAAD-funded publications in three bibliometric databases (Dimensions, Scopus and Web of Science) was compared. Then publication, collaboration and impact indicators were applied (rate of excellence or PP top 10%) on the bibliometric database with the best coverage (Web of Science).

Furthermore, a second data set was created to supplement the first data set by additional variables, which allows a differentiated analysis of the DAAD funding portfolio. For this purpose, the DAAD-funded publications in the Web of Science were linked to the DAAD database on personal funding. The linkage was done at the author level and was based on various fields: the name, email address, research field, the country of origin and destination of the grantees, as well as the funding period. Of the 33,812 DAAD-funded publications between 2010 and 2020 of the first dataset (Web of Science), 5,346 publications could be allocated to funding recipients. There are several reasons for – at a first glance – relatively small number of matches. Firstly, the DAAD database only includes individual scholarships awarded by the DAAD directly. No data is available for the extensive so-called project-related DAAD funding, in which the DAAD awards funding to institutions who then pass it to individual beneficiaries. Secondly, the DAAD database only contained personal funding from 2014 onwards. As publications usually appear a while after the start of a grant, only a small proportion of DAAD publications from 2014 could be linked.

The number of linked publications increased steadily over the years, reaching its peak in 2020 – from 165 in 2014 to 1,262 in 2020. Thirdly, the linkage was guided by high-quality criteria to exclude false assignments. This quality orientated approach also reduced the number of validated linked publications. The second data set allows for differentiated analyses of funding programs, academic status (graduate, doctoral candidate, postdoc, faculty member), and country of origin and destination (mobility).

Results

Comparison of bibliometric databases

Figure 1 shows the number of DAAD-funded publications between 2010 and 2020 for the bibliometric databases Dimensions, Scopus and Web of Science. The procedure described in the method section for identifying DAAD-funded publications was used for the Scopus and Web of Science databases. In the case of Dimensions, only the assignment made by the database provider could be used. It was therefore not possible to verify whether the US-DAAD program was excluded from the total DAAD-funded publications in the Dimensions database.

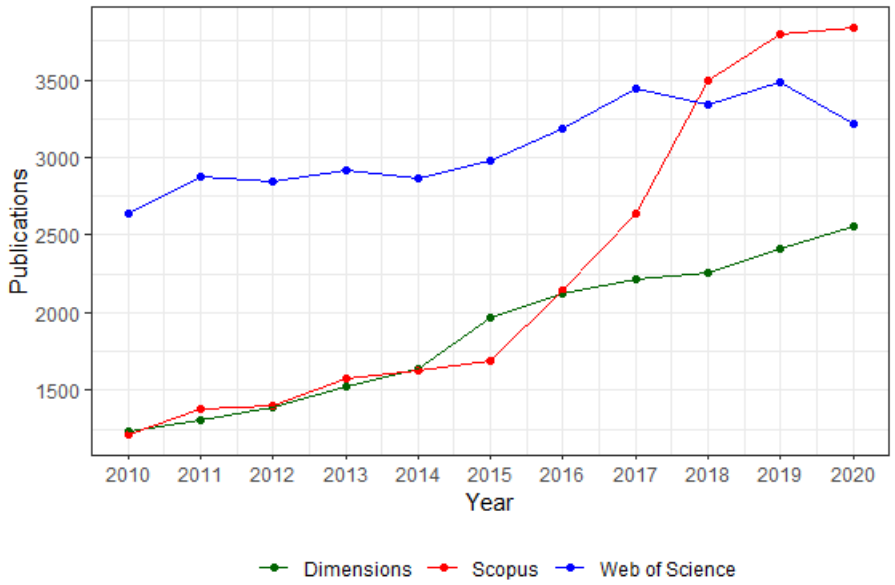


Figure 1. DAAD-funded publications in bibliometric databases (2010-2020).

The largest number of DAAD-funded publications could be identified in the Web of Science[1] (33,812), followed by Scopus (24,820) and Dimensions (20,635). The results show that the Web of Science contains the highest number of DAAD-funded publications in the period covered by the study. Scopus has caught up since 2015 and

[1] The following indexes were included in our study: Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index (AHCI), and Conference Proceedings Citation Index (CPCI).

has exceeded the Web of Science in the absolute number of DAAD-funded publications since 2018. The number of DAAD-funded publications is also increasing in Dimensions, but overall, it lags behind the two other databases.
How should the increases in Figure 1 be interpreted? Are there more and more DAAD-funded publications? The main cause of the increases is the improved coverage of funding information in bibliometric databases. In the Web of Science, funding information was initially only included in the database for certain publications (articles and reviews in journals) from the natural and life sciences. From 2015, publications from the social sciences were added, followed by the humanities from 2017, along with conference publications (Clarivate Analytics 2022: 25). Thus, the proportion of publications with funding information in the Web of Science has increased from 37% in 2010 to 56% in 2020. A similar development can be seen in Scopus. Although Scopus overtakes Web of Science in terms of the absolute number of DAAD-funded publications, the database is also somewhat larger. During the period under investigation (2010-2020), Web of Science contains 27.8 million publications, Scopus 33.6 million and Dimensions even 50.5 million. The proportion of DAAD-funded publications out of the total number of publications in the respective database (Figure 2) is higher in the Web of Science (0.11%) than in Scopus (0.10%) or Dimensions (0.04%).

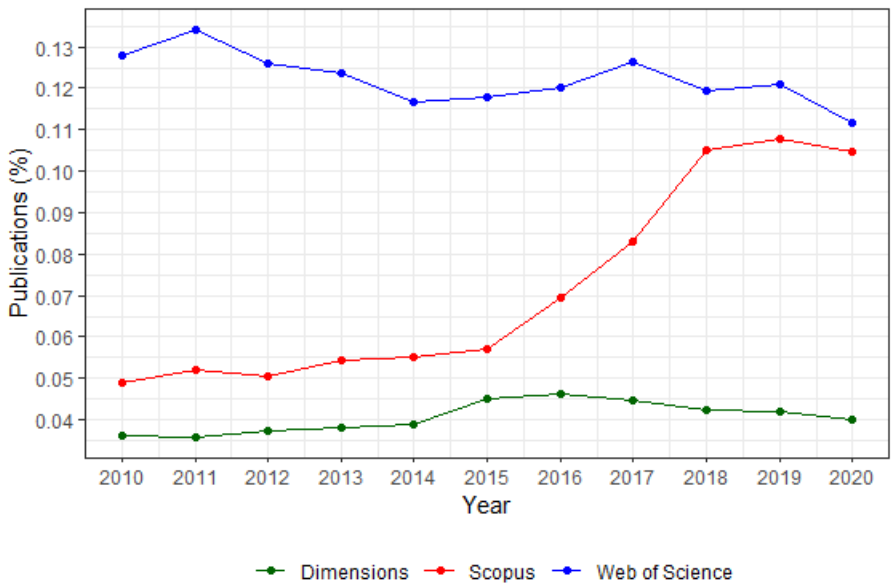


Figure 2. Proportion of DAAD-funded publications in bibliometric databases (2010-2020).

Regarding the coverage of funding information in bibliometric databases and the possibilities for further analysis, it can be concluded that the Web of Science offers the best data basis overall. For more recent analyses (from 2018 onwards), Scopus can also be used. In comparison, the Dimension database is much less suitable. Since

this study examines DAAD-funded publications from 2010 to 2020, the results in the following sections are based on the Web of Science.

DAAD-funded publications by country

According to official information from the German Academic Exchange Service, the DAAD is the largest funding organization worldwide for the international exchange of students and scholars (DAAD 2024a). Applications for funding programs do not only come from Germany, but from all over the world. Figure 3 shows the number of DAAD-funded publications per country. The assignment of a DAAD-funded publication to a country is based on the affiliations of the authors given in the publication. A DAAD-funded publication can thus be assigned not only to one, but to several countries. It is not possible to distinguish whether an author was funded by the DAAD in the respective country or worked and published with a person from that country. Figure 3 thus shows both funding and collaboration effects.

Of the DAAD-funded publications from 2010 to 2020, a total of 33,768 could be assigned to one or more countries. A total of 73,373 publication-country links were included in Figure 3, which results in an average of 2.2 countries per DAAD-funded publication.

The first thing that stands out when looking at Figure 2 is that there are hardly any white areas on the world map, i.e. there are only a few countries without a DAAD-funded publication between 2010 and 2020. The publications come from a total of 169 countries. Most DAAD-funded publications have at least one German address (27,812), followed by the USA (5,302). This means that authors in Germany are involved in 82% of all DAAD-funded publications. In particular, large countries or countries with a strong higher education and research system have numerous DAAD-funded publications. Fewer publications come from countries in Africa, Central and South America, and Asia.

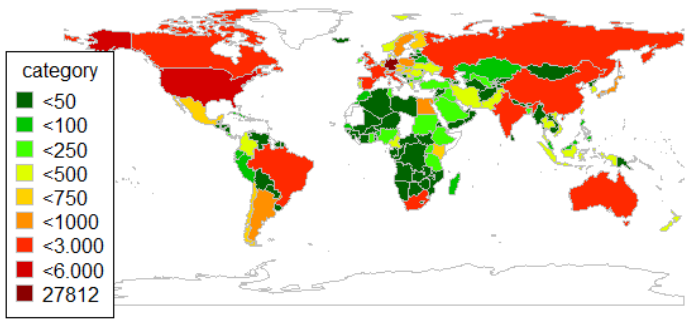


Figure 3. Number of DAAD-funded publications per country (multiple counting for international publications, 2010-2020).

While Figure 3 presents the absolute number of DAAD-funded publications, Figure 4 shows the share of these publications in relation to the country's total output. DAAD-funded publications accounted for 0.12% of total global output between 2010 and 2020. The largest number of DAAD-funded publications came from Germany, accounting for 1.6% of the total publication output from Germany. This makes the DAAD the fourth largest research funding organization in Germany after the German Research Foundation (DFG), the Federal Ministry of Education and Research (BMBF) and the Alexander von Humboldt Foundation (AvH) (see Möller 2019).

The USA, the second-largest country of DAAD-funded publications after Germany, is considerably below the global percentage (0.07% in the USA compared to 0.12% worldwide). The more than 5,000 DAAD-funded publications in which authors from the USA were involved are marginal from the perspective of the US academic system. The larger countries or countries with a strong international higher education and research system tend to have a low proportion of DAAD-funded publications. By contrast, the proportion in countries in Africa, parts of Central and South America and Asia are above the world average. Some DAAD programs are specifically aimed to support students and academics from less developed higher education systems. Figure 4 makes it apparent how important DAAD funding is, especially for countries that do not have a highly developed science and science funding system. The share of DAAD-funded publications can be seen as an indication of the importance of DAAD funding for the respective country. Despite many USA-publications, DAAD funding is less important for the USA-science system. It is much more important in Africa, parts of Central and South America and Asia.

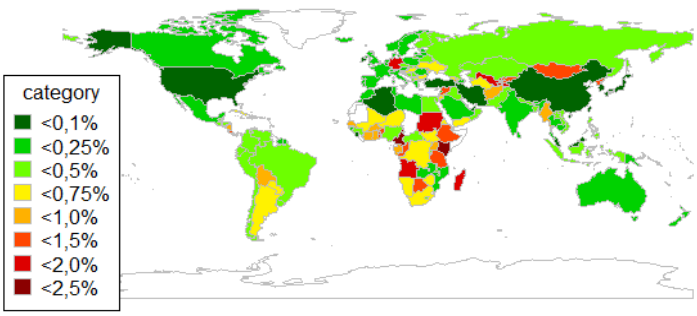


Figure 4. Proportion of DAAD-funded publications among the total publications of a country
(multiple counting for international publications, 2010-2020).

Share of international publications

Figure 5 compares the share of international publications for the DAAD, Germany and the world. Publications in which authors from more than two countries were involved are classified as international publications. Overall, the percentage of international publications has increased since 2010. However, there are some significant differences between the units of analysis: publications with at least one German address show a degree of internationalization that is more than twice as high (2020: 57%; Germany) as that of all worldwide publications (2020: 25%; World). It should be noted here that worldwide indicators are more strongly influenced by very large countries (especially the USA) and their publication output. Larger countries tend to have lower proportions of international publications because there are more national opportunities for collaboration than in smaller countries.

The publications funded by the German Academic Exchange Service (DAAD) have an international share of 75% (2020), which indicates that they have an even stronger international focus than all publications from Germany or all worldwide publications. We have differentiated the DAAD publications into those with a German affiliation and those without, reflecting the fact that the DAAD funds mobility from Germany to other countries as well as from other countries to Germany. The DAAD also finances scholarship in other countries even if the scholarship holders do not come to Germany. For DAAD publications with a German affiliation (DAAD with Deu), the internationalization share is 81% (2020), and for those without a German affiliation (DAAD without Deu), it is 46% (2020). Both percentages are considerably higher than those of the respective comparison groups (Germany and the World). The results show that the internationally oriented DAAD funding (see Figures 3 and 4 above) is not only manifested in publications in many countries. It is also underpinned by a high proportion of international collaborative co-authorships.

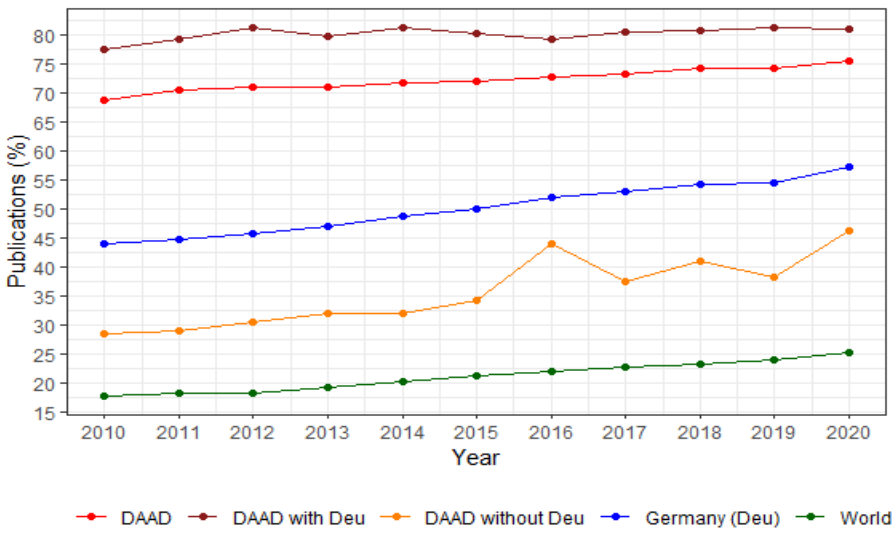


Figure 5. Share of international publications.

Excellence rate

The excellence rate (PP top 10%) presented in Figure 6 shows the percentage of DAAD-funded publications that are among the top 10%-highly cited publications worldwide. The citation indicator is calculated in two steps. First, the ten percent of journal articles and reviews with the highest citation rates are determined for each subject area and year separately. The citations are counted over a three-year period. The full-counting method is applied. Figure 6 shows that the excellence rate for worldwide publications is – as expected – 10% (world benchmark). Our calculation exactly corresponds to the 10% benchmark due to an elaborate method that uses fractionated count if more than one publication is on the PP top 10% threshold or if a publication is not top 10% highly cited in all its subject fields.

The excellence rate of Germany (Deu) was 14.1% in 2010 and 12.4% in 2020. Although the excellence rate is still higher than the global excellence rate of 10% (world benchmark)., there are various reasons for the decline: On the one hand, the data basis of the Web of Science has changed; on the other hand, the excellence rate of emerging science countries, especially China, has increased in recent years. Overall, this has led to a decline in the excellence rates of most Western European countries and of the USA (see Stephen and Stahlschmidt 2022:7).

The excellence rate of DAAD-funded publications also decreased during the period under investigation, from 12.6% to 10.2%. If we differentiate DAAD-funded publications according to whether they have a German affiliation (DAAD with Deu) or do not have a German affiliation (DAAD without Deu), we see differences in the excellence rate. The larger number of publications (DAAD with Deu) shows a similar trend as the DAAD-funded publications as a whole. DAAD-funded publications without a German affiliation initially have a higher rate of excellence, which drops sharply from 2016 onwards, falling below the global benchmark of 10%.

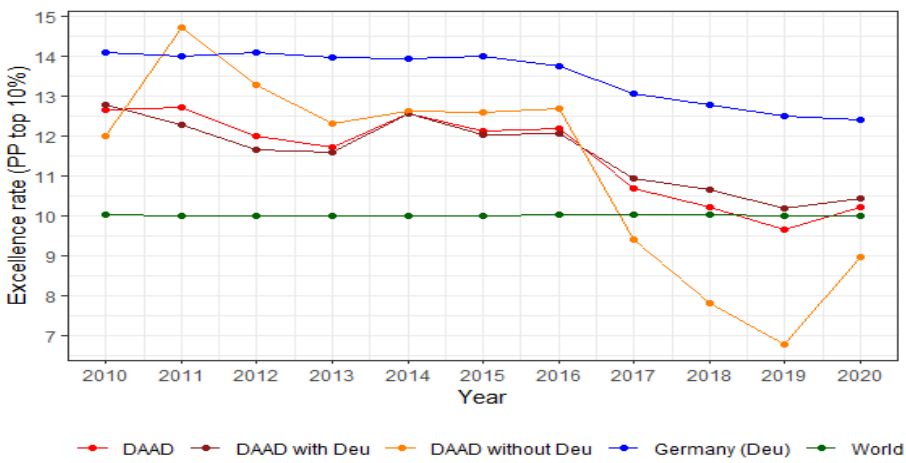


Figure 6. Excellent rate (PP top 10%) of DAAD-funded publications in comparison with Germany and the world.

The trends in the excellence rates of the DAAD-funded publications raise questions: Why is there a strong decline in the excellence rate of the DAAD-funded publications without a German affiliation? How do the different DAAD funding programs and the academic degree of the funding recipients (graduates, doctoral candidates, postdocs, faculty members), but also the country of origin and destination, affect the excellence rate?

Excellence rate by academic groups and funding programs

To answer the above questions, the bibliometric data of the DAAD-funded publications were supplemented by data from funding recipients provided by the research funder (see *Method* section). The descriptive results for various publication sets are shown in Table 1.

A total of 5,016 DAAD-funded publications were included in the analysis, given that only journal publications of the type of article and review are considered for the calculation of the excellence rate. The number of publications added up for the academic groups (row 3) and the funding program groups (row 8) is slightly higher, because some individuals received multiple funding, and it was not always possible to clearly assign the publications to a single academic group or a single funding program. In these cases, publications were assigned to multiple publication sets. First, it is noticeable that the excellence rate of the linked publications is higher (row 2, 14.3%) than that of all DAAD-funded publications (row 1, 11.5%). DAAD individual funding has a higher impact than the entire DAAD project funding. Furthermore, the excellence rate of individual funding depends on the academic status of the funding recipients (see rows 4–7). The lowest excellence rate was found among graduates (2.7%), although this group only accounts for 92 publications. Most publications were produced by doctoral students (3,153), with an excellence rate of 11.5%, and by postdocs (1,995), with an excellence rate of 19.4%. Faculty members produced 108 publications, with an excellence rate of 17.2%. In summary, it can be stated that the rate of excellence – as expected – increases with the academic degree. The faculty members are an exception with a lower excellence rate than the postdocs.

Table 1. Excellence rate of DAAD-funded publications by academic group and program.

No	Publication sets	Pub.	PP top 10%
1	DAAD-funded publications (Web of Science, 2010-2020)	31,978	11.5%
2	DAAD-funded publications linked to the DAAD scholarship database (from 2014)	5,016	14.3%
3	**Academic group**	**5,348**	
4	Graduates (Grad)	92	2.7%
5	Doctoral candidates (Doc)	3,153	11.5%
6	Postdocs (Postdoc)	1,995	19.4%
7	Faculty members (Faculty)	108	17.8%
8	**Funding program (funded group(s), origin)**	**5,581**	
9	Binationally supervised dissertations (Doc, not Deu)	132	9.7%
10	Third-country scholarships / Sur place (SPDL) (85% Doc, 14% Grad, not Deu)	240	7.3%
11	EPOS program (82% Doc, 18% Grad, not Deu)	22	4.5%
12	Research grants for dissertations (Doc, Deu)	408	15.1%
13	Research grants for dissertations in Deu (Doc, not Deu)	495	10.6%
14	Research grants - short-term (75% Doc, 23% Postdoc, not Deu)	170	15.3%
15	Research grants - long-term (85% Doc, 14% Postdoc, not Deu)	792	12.2%
16	Graduate School Scholarship Program (Doc, not Deu)	191	12.1%
17	Co-financed program (64% Doc, 31% Postdoc, not Deu)	1,202	13.6%
18	Postdoc grants from Germany to abroad (Postdoc, Deu)	915	23.0%
19	Postdoctoral Researchers International Mobility Experience (P.R.I.M.E.) (Postdoc, Deu)	265	15.1%
20	Special research programs (65% Postdoc, 35% Doc, 80% not Deu, 20% Deu)	445	13.7%
21	Other (38% Doc, 35% Postdoc, 21% Faculty, 52% not Deu, 48% Deu)	304	10.7%

A distinction according to the academic group of the funding recipients over time (not shown in Table 1) provides the following results: While in 2014, 71% of DAAD-funded publications came from the postdocs or faculty members, in 2020 this share was only 25%. During the same period, the share of publications by doctoral candidates increased from 27% to 73%. In addition, there were also changes in the countries from which the funding recipients came. In 2014, 87% of the funding recipients came from countries with an excellent rate above the global benchmark. In 2020, this was only 53%. At the same time, the excellence rate decreased from 21.1% (2014) to 11.2% (2020). These findings are relevant when interpreting the decreasing excellence rates of DAAD-funded publications shown in Figure 6. If the publication structure of individual scholarship funding corresponds to that of the total dataset of DAAD-funded publications, then the falling excellence rates for the DAAD could also be attributed to changes in the DAAD funding portfolio. A reduced funding of postdocs and faculty members in favor of doctoral students from less research-intensive countries would be a plausible explanation for the decline in the DAAD excellence rates in Figure 6.

The rows 9 to 21 show the excellence rates of the DAAD funding program groups for individual funding. The individual programs were clustered into program groups

that were established with the support of the DAAD. As the rate of excellence depends on the academic group, we show the proportion of publications by academic group in each program. Shares below 10% are omitted. We also make a broad distinction between the origin of the submitted applications (Germany (Deu) or not Germany (not Deu)).

We cannot go into all the programs in detail but concentrate on specific examples. The EPOS program (development-related postgraduate studies, row 11) has the lowest number of publications (22) and the lowest excellence rate (4.5%). The aim of the program is to qualify specialists and leaders from emerging countries as future decision-makers (DAAD, 2024b). The program is not intended to research purposes and thus the excellence rate is small and not an appropriate measurement to assess the program. Furthermore, the number of publications is too low to achieve meaningful results and only a small proportion of the funding recipients have even published anything at all. The EPOS program is a good example of why program objectives and indicators should be evaluated in relation to each other to avoid inappropriate conclusions.

The program group Postdoc Scholarships from Germany to abroad (row 18) is a different case. This category includes both short-term scholarships (three to six months) and one-year scholarships for postdoctoral researchers. The program aims to carry out (self-selected) research projects abroad (DAAD, 2024b) and is dedicated to research. Scientific publications are therefore the expected results. The Postdoc Scholarships from Germany to abroad has 915 publications and the highest excellence rate (23.0%) of all DAAD programs. In contrast to the EPOS program, the excellence rate is a suitable indicator for the program's objectives and indicates a high level of research excellence.

The program group Postdoc Scholarships from Germany to abroad has gradually been phased out in recent years. The P.R.I.M.E. program (Postdoctoral Researchers International Mobility Experience, line 19) has taken its place with a similar funding objective. It is designed for postdocs from Germany who would like to pursue an annual own research project at a research institution abroad and, after their return to Germany, receive 6 months of funding at a German university. Like its predecessor, it is therefore a funding program specifically geared towards research. The success rate of the P.R.I.M.E. program (15.1%) is among the highest of the program groups listed. However, it is lower than that of its predecessor program.

A special aspect of the two postdoc fellowship programs mentioned above is that the grantees come from Germany and move abroad. In most of the other programs, the grantees come from abroad. They either receive funding directly in their country of origin (e.g., in the case of third-country fellowships, row 10) or move to Germany for the funding period (e.g., research fellowships for doctoral studies in Germany, row 13).

The results indicate that the excellence rate of the funding programs is influenced by various factors: (a) The excellence rate depends on the funded academic group (graduates, doctoral candidates, postdocs and faculty members). (b) DAAD funding programs are not always primarily focused on research. They can also pursue other funding objectives, such as strengthening the higher education and science systems

of the Global South or sustainability aspects (cf. in Table 1 e.g., EPOS, line 11, and Third Country Scholarships & SPDL, line 10). Funding programs that are not primarily research-oriented tend to have lower excellence rates. (c) The research conditions of those funded abroad are not always comparable with those in Western countries. These affect, for example, the training and supervision of doctoral candidates as well as the financial opportunities to publish in internationally renowned journals. This can have both positive and negative effects on the excellence rate. (d) Research topics could also be country-specific or regional and influence access to international journals and perception by the global scientific community. For example, publications that deal with specific crops and the climatic conditions in a particular region may be less relevant for international journals and their global audience than other international research topics.

Modeling the factors influencing the excellence rate

The descriptive results above suggest that the excellence rate depends on various factors. In order to analyze this in more detail, we calculated various logistic regression models with the excellence rate (pp top 10%) of the scholarship holders' publications as the dependent variable. Table 2 presents the model with the independent variables academic grade, gender, and the excellence rate of the country of origin and destination. The reference category is male doctoral candidate.

Table 2. Logistic regression. Excellence rate (PP top 10%) of DAAD-funded publications by academic group, gender, country of origin and country of destination.

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.47755	0.38855	-8.950	< 2e-16	***
Graduates	-1.82520	0.71682	-2.546	0.010889	*
Postdocs	0.42550	0.08439	5.042	4.61e-07	***
Faculty members	0.36689	0.24934	1.471	0.141171	
Gender (female)	-0.20582	0.08160	-2.522	0.011658	*
Country of origin (PP top 10%)	3.83276	1.55842	2.459	0.013917	*
Country of destination (PP top 10%)	8.96088	2.71467	3.301	0.000964	***

Sig.: '***' 0.001, '**' 0.01, '*' 0.05, '.' 0.1

We see statistically significant differences in the academic group. Graduates have a lower rate of excellence than the reference group, while postdocs and faculty members have a higher rate. The result is not significant for the faculty members. Women have a significantly lower excellence rate, after controlling for the other variables. The excellence rate of a grantee's publication is also influenced by the excellence rate of the country of origin and destination. In both cases, the excellence rate increases significantly, whereby the country of destination has a greater influence than the country of origin. Grantees who come from higher education and science systems with a higher excellence rate or who move to such a system tend to achieve a higher excellence rate with their publications.

We have calculated further models (not included in this paper) in which the funding programs were included as an independent variable. The funding programs themselves had no significant influence on the excellence rate of DAAD-funded publications when the other independent variables listed above were included in the model. This result indicates that the funding programs themselves have no direct influence on the excellence rate. However, the programs do lead to certain funded academic groups from certain countries of origin and destination. This mediated influence leads to impact differences between the funding programs.

Discussion and conclusion

The present study shows that – although the DAAD does not generally ask funding recipients to acknowledge the funding source – a large number of DAAD-funded publications could be identified in bibliometric databases. The Web of Science contains more DAAD-funded publications (33,812, 2010-2020) than Scopus and Dimensions (Figures 1 and 2).

The DAAD-funded publications published between 2010 and 2020 were affiliated with institutions from 169 countries (Figure 3). 82% of these publications also contained a German affiliation. Most DAAD-funded publications have affiliations with institutions in large and/or research-intensive countries (Figure 3). These are the countries that also account for the majority of global publications. However, if we look at the share of DAAD-funded publications in the total output of each country, it becomes clear that the higher education and science systems in Africa, Central America and parts of South America and Asia benefit particularly strongly from DAAD funding (Figure 4). In addition, the DAAD-funded publications have an above-average proportion of international collaborations (affiliations from more than two countries) compared to German or worldwide publications (Figure 5).

The excellence rate of DAAD-funded publications was almost consistently above the global benchmark during the research period (Figure 5). However, as in Germany and other Western countries, the DAAD's excellence rate declined at the end of the reporting period.

Individual DAAD funding has a higher rate of excellence than overall DAAD funding. The differentiated analysis according to academic groups and programs show that the excellence rate of DAAD-funded publications increases with the academic degree (from graduates through doctoral students to postdocs). Publications by Faculty members, on the other hand, have a slightly lower excellence rate than postdocs. If the grantees come from countries with a higher excellence rate or go to such a country, the excellence rate tends to be higher.

The excellence rate is not suitable for every group of funding recipient or every funding program. The excellence rate is a suitable indicator for funding programs that are specifically geared towards research. Some programs are targeted at graduates and doctoral students from countries with less developed higher education and science systems or would like to expand or deepen the competencies of decision-makers in the domestic higher education sector. Here, the excellence rate is a less suitable measure for evaluating these programs. In the context of responsible, evaluative bibliometrics, the aim is to correlate program objectives and indicators to

appropriately assess the scope of the conclusions (see Leiden Manifesto for research metrics, Hicks et al. 2015). Nevertheless, or precisely because of this, we believe it is essential to examine the output dimension of research funding using various methods, including the bibliometric method of funding acknowledgement analysis. Bibliometric analyses provide additional insights that place the monitoring of research funders and their performance on a broader, evidence-based information foundation.

Although the DAAD does not specify how funding should be acknowledged, a high number of DAAD-funded publications were identified. However, it would be useful if not only the DAAD, but also all research funders, asked their grantees to acknowledge the funding in a specific form. For this purpose, a standard text should be provided that includes the official spelling of the research funding organization in the local language and English, the acronym and a grant number. It would be advantageous if both the name and the acronym of the research funder were unique in the international context. This also applies to the funding number. This helps to identify the publications of the respective funder and the allocation to specific projects and funding programs would be feasible.

Due to the high effort required to identify the funded publications, valid rankings of research funders are much more difficult to realize than the international university rankings mentioned in the introduction. In addition, most research funders are nationally oriented. Since national funders only select their applicants from the pool of domestic researchers, the impact of the respective country also has a high influence on the impact of the respective funder. An international ranking of research funders would thus indirectly replicate the country-specific differences.

Acknowledgments

The study was financially supported by the German Academic Exchange Service (DAAD). The bibliometric analyses using Web of Science and Scopus were carried using the infrastructure of the German Competence Network Bibliometrics (funded by the Federal Ministry of Education and Research (BMBF), funding code 16WIK2101A). We would like to thank the staff of the DAAD, especially Dr. Jan Kercher and Dr. Simone Burkhart, for their content-related and administrative support in conducting this study.

References

Álvarez-Bornstein, Belén, and Michela Montesi. 2021. “Funding Acknowledgements in Scientific Publications: A Literature Review.” *Research Evaluation* rvaa038. doi: 10.1093/reseval/rvaa038.

Clarivate Analytics. 2022. “Web of Science Core Collection. XML User Guide.”

Costas, Rodrigo, and Alfredo Yegros-Yegros. 2013. “Possibilities of Funding Acknowledgement Analysis for the Bibliometric Study of Research Funding Organizations: Case Study of the Austrian Science Fund (FWF).”

DAAD. 2024a. “Deutscher Akademischer Austauschdienst (DAAD).” *Webseite*. Retrieved July 8, 2024 (https://www.daad.de/de/).

DAAD. 2024b. *Jahresbericht 2023*. Bonn.

Hicks, Diana, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. 2015. “The Leiden Manifesto for Research Metrics.” *Nature* 520(7548): 429–31. doi: 10.1038/520429a.

Meier, Andreas, Bernhard Mittermaier, Torger Möller, Matteo Ottaviani, Barbara Scheidt, and Stephan Stahlschmidt. 2023. *Publikationen Aus DFG-Geförderten Projekten – Praxis Und Nutzbarkeit von Funding Acknowledgements*. Bonn: Deutsche Forschungsgemeinschaft.

Möller, Torger. 2019. “The Impact of Research Funding Agencies on the Research Performance of Five European Countries – A Funding Acknowledgements Analysis.” in *Proceedings of the 17th International Conference on Scientometrics & Informetrics (ISSI)*, pp. 2279–87.

Möller, Torger, Barbara Scheidt, and Andreas Meier. 2024. “Are There Factors That Influence the Quality of Funding Acknowledgements in Publications?” in *Proceedings of the 28th International Conference on Science, Technology and Innovation Indicators (STI2024)*. Berlin: STI2024. https://doi.org/10.5281/zenodo.14174157

Möller, Torger, Marion Schmidt, and Stefan Hornbostel. 2016. “Assessing the Effects of the German Excellence Initiative with Bibliometric Methods.” *Scientometrics* 109(3):2217–39. doi: 10.1007/s11192-016-2090-3.

Sirtes, Daniel. 2013. “Funding Acknowledgements for the German Research Foundation (DFG). The Dirty Data of the Web of Science Database and How to Clean It Up.” Pp. 784–95 in *Proceedings of the 14th International Society of Scientometrics and Informetrics Conference*. Vol. 1.

Stephen, Dimity, and Stephan Stahlschmidt. 2022. *Performance and Structures of the German Science System 2022*. Nr. 5-2022.

Wang, Congying, Brent Jesiek, and Wei Zhang. 2024. “Elevating International Collaboration and Academic Outcomes through Strategic Research Funding: A Bibliometric Analysis of China Scholarship Council Funded Publications.” *Scientometrics* 129(7):4329–51. doi: 10.1007/s11192-024-05054-8.

van Wijk, Erik, and Rodrigo Costas-Comesaña. 2012. *Bibliometric Study of FWF Austrian Science Fund 2001-2010/11*. Zenodo. doi: 10.5281/ZENODO.17851.

https://doi.org/10.51408/issi2025_030

Balancing Accuracy and Explainability: An Ensemble-KAN Model for Patent Grant Prediction

Jing Shi[1], Xinyi Peng[2], Xizhen Qiao[3], Ye Chen[4], Xiao Liu[5]

[1]*shijing11@smail.nju.edu.cn*

Laboratory of Data Intelligence and Interdisciplinary Innovation, School of Information Management, Nanjing University, No. 163 Xianlin Avenue, Nanjing (China)

[2]*202205571025@smail.xtu.edu.cn*

School of Automation and Electronic Information, Xiangtan University, Yanggulang, western suburb of Yuhu District, Xiangtan (China)

[3]*202205571036@ smail.xtu.edu.cn*

School of Automation and Electronic Information, Xiangtan University, Yanggulang, western suburb of Yuhu District, Xiangtan (China)

[4]*chenye@nju.edu.cn*

School of Information Management, Nanjing University, No. 163 Xianlin Avenue, Nanjing (China)

[5]*liuxiao730@xtu.edu.cn*

School of Automation and Electronic Information, Xiangtan University, Yanggulang, western suburb of Yuhu District, Xiangtan (China)

Abstract

A patent is valuable intellectual property only when granted and held for the long term, and patent grant prediction is a potential strategy for reducing the uncertainty of innovation. Existing machine learning-based prediction models lack interpretability, making it difficult to effectively mitigate innovation risks. This study proposes a novel model for patent prediction that combines high predictive accuracy with strong interpretability. (1) First, we employ the KAN model for prediction, which replaces traditional neural networks with spline functions, endowing the model with interpretability and the ability to generate formula. (2) Additionally, we introduced ensemble learning to enhance the performance of the KAN model, resulting in the development of the EN-KAN model. We tested the model on Electronic Communications datasets and demonstrated strong performance while maintaining high interpretability. EN-KAN directly generates mathematical formulas, providing a more accurate and intuitive representation of the impact of different factors on the prediction results. (3) Moreover, our study reveals that factors at the examiner-level and the patent-level have the greatest impact on patent grants.

Introduction

Patents operate on a fundamental principle of exchanging public disclosure for legal protection, offering innovators a pathway to secure exclusivity, establish technological monopolies, and generate economic returns (Nordhaus, 1969). However, the failure of a patent application to be granted can impose substantial losses on innovators, not only in terms of the time, resources, and financial investment expended but also through the unintended exposure of proprietary technologies, potentially forfeiting competitive advantages (Millar et al., 2018). Early prediction of patent grant outcomes can empower innovators by improving the likelihood of success, informing strategic decision-making in the application process, and guiding investment priorities. Although patent laws mandate that applications meet the criteria of novelty, inventiveness, and utility (Liegsalz & Wagner, 2013), these attributes are often subject to complex and multifaceted influences. The interpretive judgments of patent examiners further complicate the process, as their decisions are neither fully transparent nor easily predictable. Combined with the lengthy application cycles and extensive documentation requirements, these challenges make early prediction of patent grant outcomes a complex and urgent challenge.

To address this challenge, prior research has explored various approaches, including traditional statistical methods and heuristic analyses, to predict patent grant probability (Drivas & Kaplanis, 2020; Gans et al., 2008; D. Yang, 2008; Yao & Ni, 2023). However, these methods often suffer from limitations, such as oversimplification of complex interactions among influencing factors. Machine learning (ML) approaches, which can extract latent patterns from large-scale empirical data, have increasingly been employed to tackle this problem. For instance, ML models have been used to predict the likelihood of innovation failure by identifying significant predictors within voluminous datasets (Yao & Ni, 2023). Despite their promising predictive accuracy, the inherent “black box” nature of most ML algorithms has raised concerns regarding their interpretability, leading to skepticism about their conclusions. This lack of transparency has hindered the dissemination and practical application of ML-based findings. While some researchers have sought to enhance interpretability by appending post hoc explanation models, such methods often yield explanations that are either overly generalized or insufficiently specific to the contexts of patent examinations. Furthermore, prior studies have highlighted the variability in patent grant outcomes across different patent authorities and technological fields (Alcácer et al., 2009), emphasizing that influencing factors are not universally consistent but contingent on the specific jurisdiction and field of innovation. How these contextual factors influence patent grants remains unclear.

This study proposes a novel interpretable machine learning model, Ensemble Kolmogorov-Arnold Network (EN-KAN), to investigate the factors influencing the early prediction of patent grant. This model is designed to achieve two primary research objectives. First, unlike conventional ML models that rely on post hoc interpretability enhancements, KAN incorporates interpretability as a core feature of its design, employing knowledge embeddings and structured influence analysis (Liu et al., 2024). By comparing KAN with several benchmark algorithms, we demonstrate its efficacy and provide visualized explanations of its findings. Our results identify critical predictors of patent grant success elucidating their underlying mechanisms by formula. Second, we examine the differential impacts of patent examination authorities, uncovering jurisdiction-specific patterns and highlighting the role of institutional and procedural variations in shaping grant.

The contributions of this study are twofold. First, we introduce a self-explanatory model that accurately predicts patent grant probabilities while identifying key determinants of patent success. By integrating interpretable methodologies, this research advances the understanding of patent grant processes and provides a robust framework for examining the drivers of patent approval. Second, this study offers comparative insights across diverse technological domains and patent jurisdictions, addressing gaps in the literature regarding the contextual variability of influencing factors. These findings have practical implications for both patent applicants and examiners. For innovators, the results offer actionable guidance for crafting application strategies to maximize the probability of success and minimize uncertainties, ultimately enhancing the commercial value of patents. For patent examiners, the insights enable optimization of examination workflows, improving efficiency by focusing on the most impactful variables. Through these contributions, this research not only advances academic discourse but also supports evidence-based decision-making in the patent ecosystem.

Literature review

The influencing factors of patent grant

The factors influencing patent grant can be categorized into five levels: patent, application, applicant and inventor, examiner, and other factors. Table 1 provides a summary of these levels and their corresponding factors.

Patent Level focuses on the intrinsic characteristics of the innovation, including novelty, innovativeness, and utility. Novelty and innovativeness are fundamental traits of patents and serve as key drivers of technological breakthroughs, playing a decisive role in patent grant. Prior studies have employed various measures to assess novelty, such as the number of International Patent Classification (IPC) categories

involved (Harhoff & Wagner, 2009; Liegsalz & Wagner, 2013), the number of references cited (G. Yang et al., 2023), and the Herfindahl index (a measure of concentration) of cited patent classes. Emerging research highlights the role of scientific knowledge in technological innovation, finding that patents utilizing more scientific knowledge exhibit higher innovativeness (C. Lee et al., 2018). Utility reflects the practical applicability or industrial use of an invention. A common metric for utility is the generality index, which measures the breadth of subsequent inventions benefiting from the patent (Niosi, 2006). Public procurement patents tend to have higher generality (Raiteri, 2018) and patents with greater generality demonstrate sustained competitiveness (P.-C. Lee, 2021).

Application Level emphasizes the quality of the application documents, including indicators such as the number of pages, titles, abstracts, claims, and the length of claims. Claims delineate the scope of the patent. While a higher number or broader scope of claims increases examination complexity and may prolong the review process (Liegsalz & Wagner, 2013), research also suggests a positive relationship between the number of claims and patent grant. A patent with numerous independent claims is perceived as robust in legal terms (Harhoff & Wagner, 2009; Y.-G. Lee & Lee, 2010). The word count of the first claim is another commonly used indicator, reflecting the patent's protection scope (Sampat & Williams, 2019). Moreover, particular attention is given to Patent Cooperation Treaty (PCT) applications. PCT filings, which enable the extension of patent protection to multiple countries while minimizing costs and complexities, positively impact patent grant rates (Harhoff & Wagner, 2009)

Applicant and inventor level explores the influence of applicant and inventor characteristics, such as quantity, nationality, and historical experience. Analysis of USPTO data reveals that U.S. nationality increases the likelihood of patent approval, whether as applicants or inventors (Drivas & Kaplanis, 2020). Some patent office's exhibit preferential treatment toward domestic applicants (D. Yang, 2008), leading to higher granting probabilities for local inventors. Additionally, in areas of technological specialization, domestic inventors show stronger positive effects (Webster et al., 2014). However, excessive domestic collaboration may reduce the probability of patent grants. In contrast, international collaborations tend to confer advantages (Guellec & de la Potterie, 2000). Applicants with prior success in securing patents are more likely to achieve subsequent grants (Liegsalz & Wagner, 2013). Persistent efforts in filing patents also significantly enhance granting probabilities (Drivas & Kaplanis, 2020).

Examiner Level addresses the role of patent offices and examiners. Decisions on patent grant are heavily influenced by individual examiners (Lemley & Sampat, 2012), and examiner biases can distort patent allocation. For instance, examiners

may be less likely to grant patents to inventors outside their social group (Desai, 2019). They also demonstrate a tendency to approve patents for applicants of the same gender (Shen & Zingg, n.d.). Examiners' behaviors are influenced by their peers, particularly when in close physical proximity (Frakes & Wasserman, 2021). These dynamics underscore the subjective aspects of the patent examination process.

Other Factors. Additional factors include the technological field, patent application strategies, and the number of related patent filings. Comparative analyses of 30 technological fields reveal significant differences in patent review durations across domains (Liegsalz & Wagner, 2013). A Difference-in-Differences (DID) analysis by Bekkers demonstrated that increased awareness of earlier related technologies among examiners reduces patent grant probabilities (Bekkers et al., 2020).

Table 1. The relevant influencing factors of patent grant.

Dimension	*Factors*	*Sources*
Patent level	Novelty	Harhoff & Wagner, 2009; Liegsalz & Wagner, 2013; C. Lee et al., 2018; G. Yang et al., 2023
	Utility	Niosi, 2006; Raiteri, 2018; P.-C. Lee, 2021
Application level	the number of pages of application file	Yao & Ni, 2023
	the number of claims	Harhoff & Wagner, 2009; Y.-G. Lee & Lee, 2010; Liegsalz & Wagner, 2013; Marco et al., 2019
	the word count of title	Yao & Ni, 2023
	the word count of abstract	Yao & Ni, 2023
	the word count of claims	Marco et al., 2019; Sampat & Williams, 2019
	whether submit PCT application or not	Harhoff & Wagner, 2009
Applicant & inventor level	whether local applicant/inventor or not	D. Yang, 2008; Guellec & de la Potterie, 2000; Drivas & Kaplanis, 2020

	the number of applicants/inventors	C. Lee et al., 2018; Yao & Ni, 2023
	applicant's experience	Harhoff & Wagner, 2009; Liegsalz & Wagner, 2013
	the nationality of applicant	D. Yang, 2008; Webster et al., 2014; Drivas & Kaplanis, 2020
Examiner level	Examiner	Lemley & Sampat, 2012; Desai, 2019; Shen & Zingg, n.d.; Frakes & Wasserman, 2021
	the country of prior right	Guellec & de la Potterie, 2000; Yao & Ni, 2023
	The duration of examine	Harhoff & Wagner, 2009
Others	technological field	Guellec & de la Potterie, 2000; Liegsalz & Wagner, 2013
	the strategy of application	Guellec & de la Potterie, 2000
	the number of relevant applications	Bekkers et al., 2020

Interpretable Machine Learning Research

Interpretable Machine Learning (IML) seeks to provide insights into machine learning models that are understandable to humans. IML encompasses understanding data, the internal structures of models, and interpreting the results produced by these models (Allen et al., 2024; Lipton, 2018). The applications of IML span various stages of the machine learning pipeline, including the explanation of input data, the elucidation of model mechanisms, and the interpretation of output outcomes. Explanation techniques in IML can be categorized along three dimensions: intrinsic interpretability versus post-hoc interpretability, model-specific explanations versus model-agnostic explanations, and global explanations versus local explanations.

Intrinsic Interpretability vs. Post-hoc Interpretability. Intrinsic interpretability refers to the inherent transparency of a model, allowing users to understand its behavior directly through the training process. Examples of intrinsically interpretable models include decision trees (Costa & Pedreira, 2023), additive models (Agarwal et al., 2021), and models enhanced with regularization techniques such as sparsity (Hoefler et al., 2021) or smoothness (Crawshaw et al., 2022), which naturally provide high

levels of interpretability (Rudin, 2019). Recent advancements have further improved the intrinsic interpretability of deep neural networks by integrating prototypes or specific interpretability constraints into their final layers (Dong et al., 2017). In contrast, post-hoc interpretability involves applying additional methods to interpret the model or its outputs after the training phase. These methods include feature importance scoring based on backpropagation and Local Interpretable Model-agnostic Explanations (LIME) (Molnar, 2020). LIME, for example, constructs simplified surrogate models around specific input points to approximate the behavior of complex models, making it applicable to various pre-trained models and providing additional insights into their decision-making processes (Molnar, 2020).

Model-specific Explanations vs. Model-agnostic Explanations. Model-specific explanation methods are designed for types of models and do not generalize well across different model architectures. Examples include regression coefficients in generalized linear models (Rong & Bao-Wen, 2018), feature importance scores in tree-based models (Zhou & Liu, 2021), and techniques such as backpropagation or layer-wise relevance propagation in deep learning (Zhou & Liu, 2021). Conversely, model-agnostic explanation methods are applicable to a wide range of model types, offering a unified framework for interpretation. Common model-agnostic methods include Shapley values (Fryer et al., 2021),feature permutation (Covert et al., 2021), feature masking (J. Dai et al., 2015), and LIME (Molnar, 2020), which provide consistent explanatory effects across different models. It is important to note that model-specific explanation methods do not necessarily provide intrinsic interpretability. For instance, feature importance scores in decision trees and feature attribution via backpropagation are model-specific yet fall under post-hoc interpretability. Most model-agnostic explanation methods are inherently post-hoc in nature.

Global Explanations vs. Local Explanations. Global explanations aim to reveal the overall structure of the model and the general importance of all features. Examples include coefficients in linear or additive models, feature importance scores in tree-based models, and global feature attribution methods, which reflect each feature's role in the model's overall predictions. On the other hand, local explanations focus on specific inputs or subsets of inputs, providing targeted interpretations. For example, LIME and saliency map methods concentrate on individual test instances or the significant features of specific observations (Ribeiro et al., 2016). In unsupervised learning, local embedding methods such as t-SNE (t-distributed Stochastic Neighbor Embedding) (Van der Maaten & Hinton, 2008) and UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018) analyze data patterns and relationships within specific neighborhoods to explain local data.

Despite significant advancements in enhancing model transparency, current IML

approaches exhibit several limitations. Firstly, there is considerable technical heterogeneity among existing methods, with each approach typically catering to specific interpretative needs and lacking generalizability. This fragmentation leads to inconsistent explanatory outcomes across different methods, thereby complicating users' understanding of model behavior. For instance, some methods emphasize global feature importance while others focus on local instance explanations; employing multiple methods simultaneously may yield conflicting conclusions. Additionally, varying assumptions and focal points among different methods result in a lack of unified evaluation standards, undermining the reliability and consistency of explanations. Such inconsistencies not only increase the difficulty for users to comprehend and trust the models but also risk misleading decision-making processes, thereby reducing the practical effectiveness of interpretability techniques. Consequently, there is an urgent need to develop more unified and coordinated interpretability frameworks to mitigate methodological discrepancies, enhance the consistency of explanatory outcomes, and bolster user trust.

Methodology

Data collection

We select patents in the fields of Electronic Communications (EC) for empirical analysis and comparison due to their pivotal roles in driving technological progress and economic growth. EC, as a mature and highly competitive sector, presents unique challenges in balancing innovation with the standardization of technologies. Invention patents are selected for analysis due to their emphasis on groundbreaking innovations and their rigorous examination standards. Invention patents are emphasized because they represent substantive technological innovations and generally possess higher overall market value. Moreover, the examination process for invention patents is more rigorous, with clearer and more consistent decision-making criteria, making them more predictable. Finally, invention patents offer higher data quality and richer textual information, making them particularly well-suited for training patent grant prediction models. The process of obtaining an invention patent typically involves several key stages, beginning with the filing of a patent application. After filing, the application undergoes a formal examination and the substantive examination phases. If approved, the patent is granted and published, providing the inventor with exclusive rights to the invention, typically having a protection period of up to 20 years.

The patent examination process generally spans 2 to 5 years, with an average duration of approximately 4 years, supporting the selection of a five-year observation window. Thus, invention patent applications in 2017 of the EC fields are chosen,

enabling an evaluation of whether these patents were successfully granted within 5 years. Our patent data are collected from PATSTAT (Worldwide Patent Statistical Database) and the final dataset contains 299,912 patent applications (137,257 patents are granted).

Influencing factors extraction and description

The grant status of a patent is operationalized as a binary variable, where granted patents are assigned a value of 1, and non-granted patents are assigned a value of 0. This study selects patent features as influencing factors at five levels, and the final factors and measurement methods are detailed in Table 2.

Table 2. Influencing factors selected.

Dimension	*Factors*	*Measurement*
Patent level	backward_citation	The number of backward citations.
	family_size	The family size of focal patent.
	nb_claims	The number of claims.
	nb_title_char	The word count of patent applications' title.
	nb_abstr_char	The word count of patent applications' abstract.
	is_PCT	Whether the patent is filed as a PCT application: 1 for Yes, 0 for No.
Applicant & inventor level	nb_inventors	The number of inventors.
	nb_applicants	The number of applicants.
	nb_applications	The total number of patent applications of all applicant and inventors of focal patent in 2017.
	ratio_granted	The granting rate of the applicant's patent applications in 2016.
	ctry_first_applicant	The nationality of the first applicant.
	nb_local_applicant	The number of local applicants.
	nb_foreign_applicant	The number of foreign applicants.
	nb_local_inventor	The number of local inventors.
	nb_foreign_inventor	The number of foreign inventors.
	appln_auth	The examination authority of the focal patent.
Examiner level	int_phase	Whether the patent entered the international phase: Y = 1; N = 0.
	reg_phase	Whether the patent entered the regional

		phase: Y = 1; N = 0.
	nat_phase	Whether the patent entered the national phase: Y = 1; N = 0.
	duration	The number of years from the initial patent application filing to the final decision.
Others	tech_field	The 3_digit IPC code which focal patent belongs to.
	nace_code	The NACE[1] code of focal patent.
	nb_relevant_patent	The number of relevant applications[2].

Table 3. The patent features' description.

Factors	*Mean*	*SD*	*Factors*	*Mean*	*SD*
backward_citation	7.68	38.39	nb_foreign_applicant	0.90	0.58
family_size	3.84	4.79	nb_local_inventor	0.45	1.29
nb_claims	13.26	37.4	nb_foreign_inventor	2.39	2.17
nb_title_char	8.5	4.24	appln_auth	NA	NA
nb_abstr_char	134.64	52.41	int_phase	0.37	0.48
is_PCT	0.26	0.44	reg_phase	0.08	0.27
nb_inventors	2.78	2.12	nat_phase	0.81	0.39
nb_applicants	1.08	0.48	duration	1.73	1.11
nb_applications	1095.9	1956.03	tech_field	NA	NA
ratio_granted	0.47	0.35	nace_code	NA	NA
ctry_first_applicant	NA	NA	nb_relevant_patent	0	0.03
nb_local_applicant	0.20	0.47			

Model construction

This paper proposes an ensemble learning approach based on the ENsemble Kolmogorov-Arnold Network (EN-KAN) for predicting patent grant outcomes. The proposed method enhances prediction accuracy and model generalization through systematic data preprocessing, the design and training of the KAN model, and the implementation of an ensemble learning strategy.

[1] NACE: Statistical Classification of Economic Activities in the European Community is the statistical classification system of economic activities in the European Union (EU).

[2] Technical relations are "priority-like" relations between applications which have been detected by EPO examiners, but which have not been published by a patent office.

(a) Base model

The Ensemble-KAN utilizes the Kolmogorov-Arnold Network (KAN) as the foundational model for patent grant prediction. KANs, based on the Kolmogorov-Arnold theorem, are emerging machine learning architectures recognized as powerful alternatives to multilayer perceptrons (MLPs). The KAN network exhibits significant advantages over traditional MLPs in several key aspects, particularly in weight parameter representation and function approximation methods.

According to the Kolmogorov-Arnold theorem, for any continuous multivariate real function $f:[0,1]^n \to R$, there exists a set of univariate continuous functions $\{\phi_k\}$ and $\{\psi_{k,i}\}$ such that f can be expressed as a finite nested and summative form:

$$f(x_1, x_2, \dots, x_n) = \sum_{k=1}^{2n+1} \phi_k \left(\sum_{i=1}^{n} \psi_{k,i}(x_i) \right).$$

This theorem theoretically demonstrates that multivariate continuous functions can be decomposed into a weighted sum of univariate nonlinear functions. Unlike traditional MLPs, which employ fully connected linear transformations combined with fixed activation functions, KAN networks represent each channel with learnable univariate nonlinear functions. This alignment with the Kolmogorov-Arnold decomposition enhances the function representation's conformity to the theorem's decomposition principle.

Specifically, the KAN aims to approximate a target function $f(\mathbf{x}) = f(x_1, \dots, x_n)$ as:

$$\hat{f}(x) = \sum_{k=1}^{K} g_k \left(\sum_{i=1}^{n} h_{k,i}(x_i) \right),$$

where $g_k(\cdot)$ and $h_{k,i}(\cdot)$ are learnable univariate nonlinear functions. To enhance the function space's representation capability, KAN networks incorporate learnable B-splines as the base functions, parameterizing both $h_{k,i}$ and g_k. For example, the B-spline basis functions for $h_{k,i}$ are expressed as:

$$h_{k,i}(x_i) = \sum_{j=1}^{J} \alpha_{k,i,j} B_j(x_i).$$

Similarly, for $g_k(u)$:

$$g_k(u) = \sum_{j=1}^{J'} \beta_{k,j} B_j(u),$$

where $\{\alpha_{k,i,j}\}$ and $\{\beta_{k,j}\}$ are trainable parameters. The incorporation of learnable

B-spline activation functions allows the model to adaptively adjust the univariate nonlinear mappings during training, thereby shaping the function forms according to the data distribution characteristics and enhancing the model's ability to capture complex data patterns.

Furthermore, the univariate learnable nonlinear function structure of the KAN network improves model interpretability. Since the function is explicitly decomposed into a finite sum of univariate nonlinear functions, it facilitates the analysis of input variables' individual contributions to the output, providing more intuitive explanations for the decision-making process in the task.

(b) ENsemble Kolmogorov-Arnold Network

In this paper, we propose an *ENsemble Kolmogorov-Arnold Network (EN-KAN)*, by centrally training multiple Kolmogorov-Arnold Network (KAN) models and generating combined prediction results. EN-KAN mitigates individual model biases, significantly enhancing the overall model's generalization capability. The core idea is to leverage the diversity of multiple independently trained KAN models and integrate their predictions through an ensemble decision mechanism to achieve more robust and accurate classification performance.

Figure 1 illustrates the structure of the proposed EN-KAN. The process begins with the data preprocessing stage, which includes three main steps: Data Cleaning, Normalization, and Feature Selection. These steps work together to produce a high-quality training dataset. Once the data is preprocessed, it is fed into the EN-KAN module. This module is made up of several KAN. Each KAN network starts by fitting an explainable spline function to capture the nonlinear patterns in the data. After fitting the spline functions, they are combined to form a complete KAN network. During the prediction phase, each individual KAN network makes its own prediction based on the input data. These predictions are then collected through a voting mechanism, where each KAN network casts a vote for its predicted outcome. Finally, the EN-KAN algorithm uses a Model Ensemble process to merge all the votes from the KAN networks, resulting in the final output. This structure not only enhances the prediction accuracy of the model but also maintains the interpretability of the results.

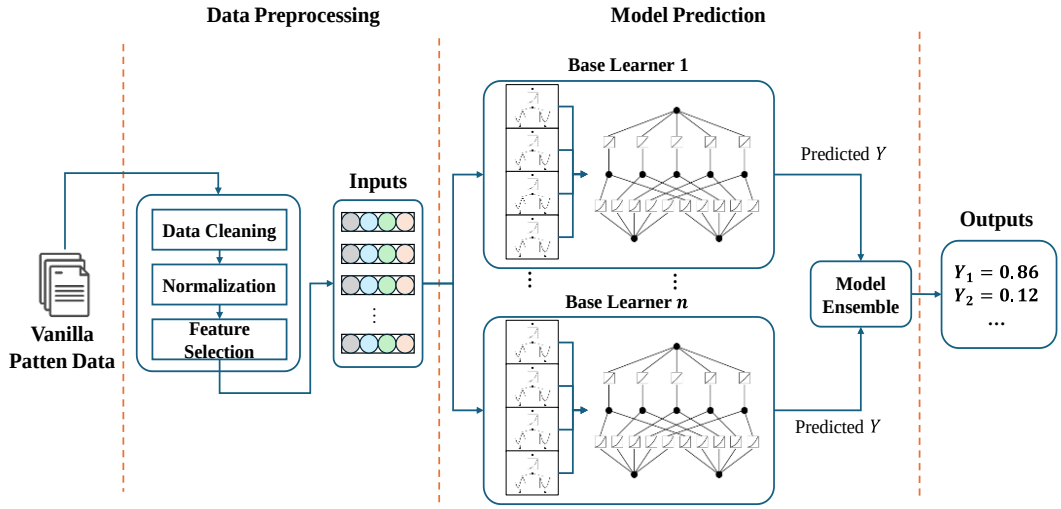


Figure 1. A high-level structure of the proposed EN-KAN.

Specifically, let there be M independent KAN models, each model m characterized by a unique parameter set θ_m. Due to different initializations and the stochastic nature of the training process, the parameter sets θ_m exhibit diversity, which is crucial for the ensemble method to improve generalization.

Formally, for each sample $\mathbf{x_i} \in \mathbf{X}_{\text{ts}}$ in the test dataset, each KAN model m generates a prediction probability vector $\hat{\mathbf{y}}_i^{(m)}$ as follows:

$$\hat{\mathbf{y}}_i^{(m)} = f_m(\mathbf{x_i}; \theta_m),$$

were, $\hat{\mathbf{y}}_i^{(m)}$ represents the predicted probabilities of sample $\mathbf{x_i}$ belonging to each class by model m. For each model m, the predicted class label $\hat{y}_i^{(m)}$ is determined by selecting the class with the highest probability:

$$\hat{y}_i^{(m)} = \arg\max_c \left(\hat{\mathbf{y}}_i^{(m)}\right)_c,$$

where c denotes the class index. The final ensemble prediction label $\hat{\mathbf{y}}_i^{(\text{ensemble})}$ is obtained by majority voting among all M models:

$$\hat{\mathbf{y}}_i^{(ensemble)} = \text{mode}\left(\{\hat{\mathbf{y}}_i^{(m)}\}_{m=1}^{M}\right).$$

The mode function returns the class that appears most frequently among the predictions of the individual models. By integrating multiple diverse KAN models,

Ensemble-KAN (E-KAN) effectively reduces the risk of overfitting inherent in single models, thereby enhancing the system's overall generalization capability.

(c) Model Training

Ensemble-KAN (E-KAN) optimizes multiple KAN networks collectively. The overall training loss $L_{\text{E-KAN}}$ is defined as the sum of the loss functions of all M models:

$$L_{\text{E-KAN}} = \sum_{m=1}^{M} \mathcal{L}_m,$$

where $\mathcal{L}_m$ represents the loss function of the m-th KAN model, defined as:

$$\mathcal{L}_m = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{\mathbf{y}}_i^{(m)}\right) + (1-y_i)\log\left(1-\hat{\mathbf{y}}_i^{(m)}\right)\right],$$

were, N is the number of samples in the training set, y_i is the true label of the i-th sample, and $\hat{\mathbf{y}}_i^{(m)}$ is the predicted probability by the m-th KAN model for the i-th sample. Thus, the overall training loss can be expressed as:

$$L_{\text{E-KAN}} = -\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{\mathbf{y}}_i^{(m)}\right) + (1-y_i)\log\left(1-\hat{\mathbf{y}}_i^{(m)}\right)\right]\right).$$

The objective is to minimize the overall training loss $L_{\text{E-KAN}}$. By optimizing multiple Kolmogorov-Arnold networks simultaneously and employing an ensemble decision mechanism, Ensemble-KAN (E-KAN) effectively enhances model performance and generalization in patent grant prediction tasks, offering a robust and efficient solution.

Result

Prediction results

Table 5 presents a comparison of the performance of our model with other models. The primary evaluation metrics include precision (P), recall (R), and F1-score. Overall, the EN-KAN, Random Forest, and KNN models demonstrated better performance compared to traditional models. EN-KAN model showed best performance, with F1-score 0.89.

Table 5. Results of different models.

Model	P (%)	R (%)	F1 (%)
EN-KAN	0.8946	0.8975	**0.8949**
RandForest	0.8643	0.8640	**0.8638**
KNN	0.8547	0.8546	0.8544
LASSO	0.7125	0.7125	0.7117
Logistics	0.7921	0.7904	0.7894

In the Figure 2, the orange curve represents the ROC curve of the EN-KAN model. The EN-KAN, Random Forest, and KNN models showed strong performance, while the LASSO model performed the worst. The Random Forest model, through the integration of multiple decision trees, effectively handles noise and feature correlations within the data, achieving performance comparable to the EN-KAN model on the dataset. However, the high performance of Random Forest comes at the cost of interpretability, as its results are often considered a "black box." In contrast, EN-KAN strikes an optimal balance between predictive performance and interpretability, making it a more suitable choice for applications requiring both robust predictions and explainable outcomes.

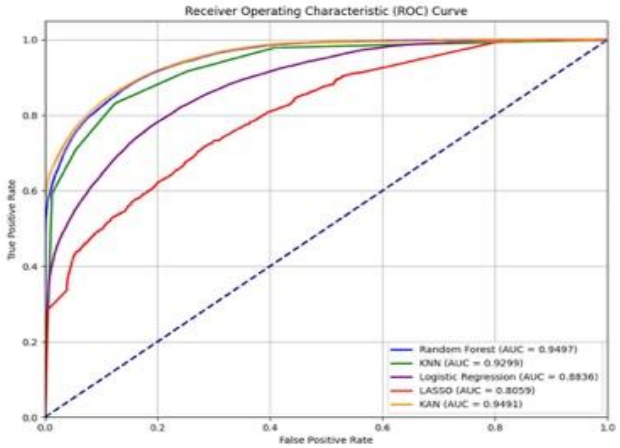


Figure 2. The ROC curve.

Figure 3 illustrates the trade-off between model interpretability and predictive accuracy, helping us understand the relative positions of different machine learning models along these two dimensions. The x-axis represents model interpretability, with models positioned further to the right being more understandable to humans. The y-axis indicates predictive accuracy, with higher positions corresponding to

better performance on the patent grant prediction task. Models in the lower right red circle are intrinsically interpretable but demonstrate lower predictive accuracy. Models in the upper left blue region achieve higher accuracy but require post-hoc interpretation methods such as SHAP and LIME to explain their predictions (Lundberg & Lee, 2017; Ribeiro et al., 2016). In contrast, models in the upper right black region—including the EN-KAN proposed in this study and its base model KAN—represent a class of neural network architectures that combine high interpretability with strong performance. These models are inherently interpretable and do not rely on external tools for post-hoc explanations. Among them, KAN provides the most transparent model structure, although its predictive performance is slightly lower than that of Random Forest. After incorporating ensemble learning, EN-KAN not only surpasses RF in accuracy but also offers superior interpretability compared to other models. The green dashed line in the figure denotes the signal-to-noise ratio (SNR), with higher values indicating that the model can more effectively capture underlying patterns, leading to improved accuracy. The transition from models in the red region to those in the blue region reflects the evolution from traditional statistical models to high-performance nonlinear models. While increased SNR supports the performance of such complex models, it often comes at the cost of reduced interpretability. The EN-KAN model introduced in this study seeks to break this trade-off by achieving an optimal balance between interpretability and predictive power.

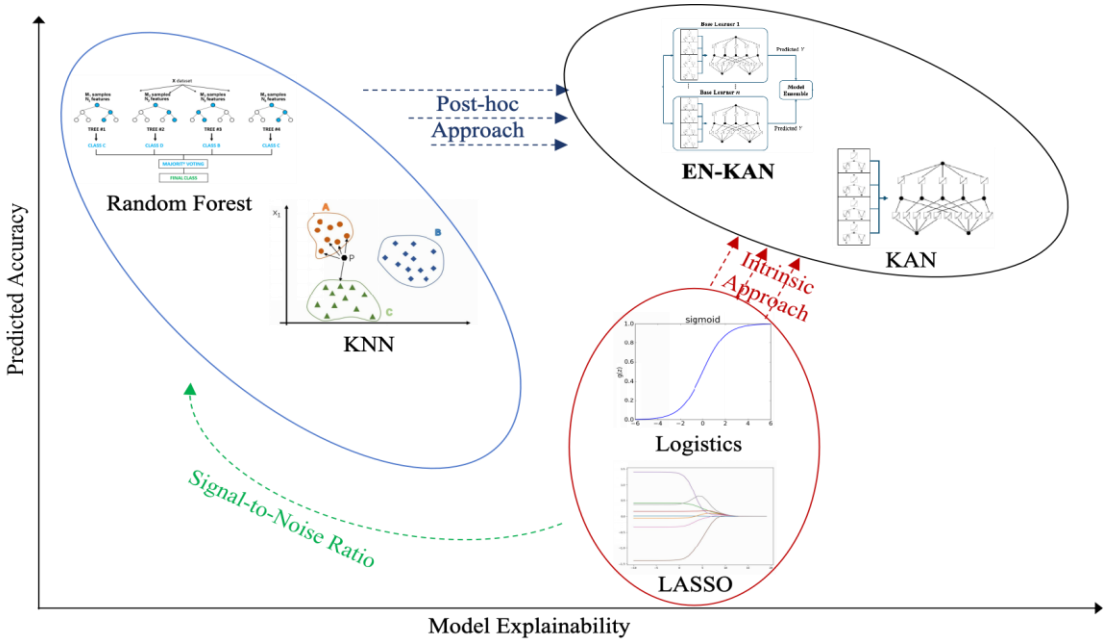


Figure 3. Explainability and predicted accuracy of different models.

Which patents are granted?

For patents in the EC field, the most significant factors are *nb_claims*, *nat_phase*, and *int_phase*. Similar to the AI field, the number of claims is the most impactful factor among all, far surpassing others. However, a key difference lies in the substantial influence of different examination phases on EC patent approvals. This may be related to the stronger global nature of EC technologies. For innovators in the EC field, participating in international patent examination procedures not only enhances the global competitiveness of their technologies but also reduces the risk of infringement by meeting international examination standards. Moreover, the examination processes at various stages are more standardized and systematic, making them critical determinants of patent approval.

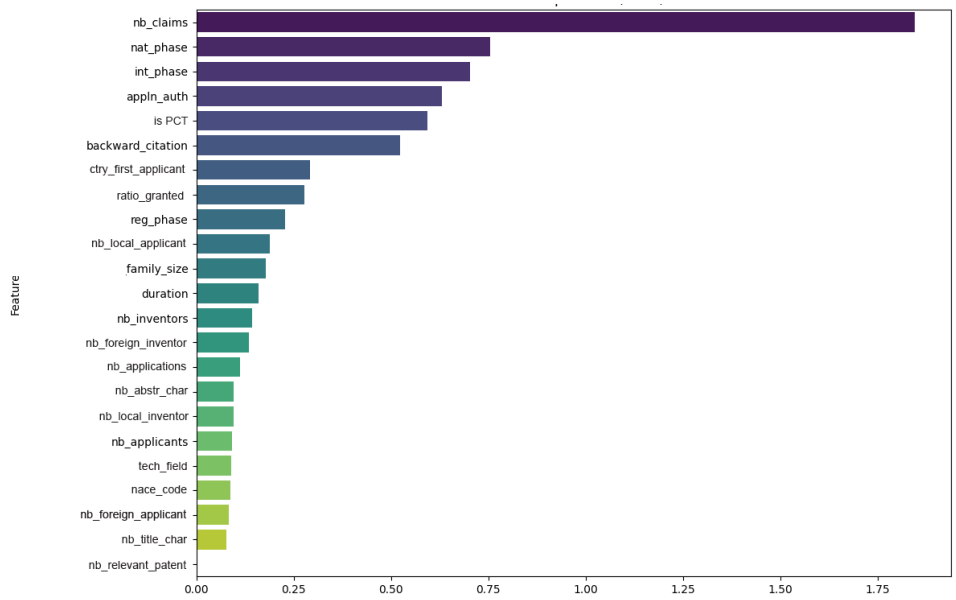


Figure 3. Feature Importance Analysis for EC Patent Grants.

Specifically, the influence of different factors on patent grants varies. First, filing a PCT application, entering the international phase, and having a higher number of claims are positive indicators of patent grants. Second, it is observed that for EC patents, a larger number of local applicants and inventors is more favorable for patent grant. Local innovators are likely to have a better understanding of the local market and regulatory environment, enabling them to submit patent applications that align more closely with examination requirements. Moreover, the involvement of local inventors may signify the practical feasibility and localized value of the technological innovation, thereby garnering greater recognition. Interestingly, unlike the other two fields, EC patent grants appear to be unrelated to backward citation. A

possible explanation is that the EC field is characterized by mature technologies with rapid innovation cycles. Innovations in this domain are often driven by new application scenarios or cross-disciplinary integration, rather than heavy reliance on existing technological foundations. Consequently, examination authorities may focus more on the practical utility of the patent rather than its connections to prior technologies.

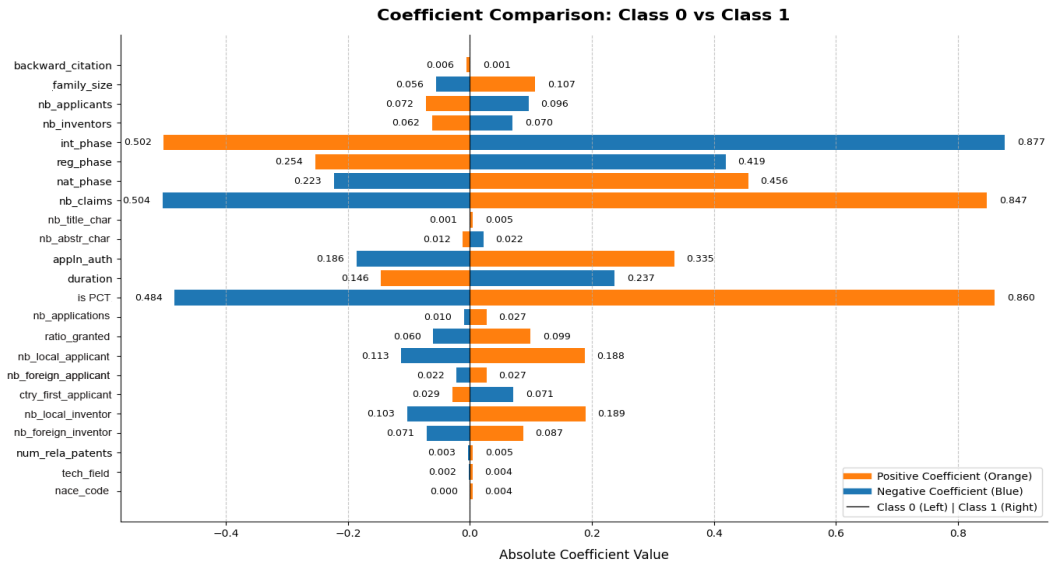


Figure 4. The coefficient comparison of influencing factors.

$$f(not\ granted) = -0.504 * nb_{claims} + 0.502 * int_{phase} - 0.484 * is_{PCT} + 0.254 * reg_{phase} - 0.223 * nat_{phase} - 0.186 * appln_{auth} + 0.146 * duration - 0.113 * nb_{local_{applicant}} - 0.103 * nb_{local_{inventor}} + 0.072 * nb_{applicants} - 0.071 * nb_{foreign_{inventor}} + 0.062 * nb_{inventors} - 0.060 * ratio_{granted} - 0.056 * family_{size} + 0.029 * ctry_{first_{applicant}} - 0.022 * nb_{foreign_{applicant}} + 0.012 * nb_{abstr_{char}} - 0.010 * nb_{applications} + 0.006 * backward_{citation} - 0.003 * nb_{relevant_{patent}} - 0.002 * tech_{field} - 0.001 * nb_{title_{char}} + 0.000 * nace_{code} + 0.776$$

Formula 1

$$f(granted) = -0.877 * \mathrm{int_{phase}} + 0.860 * \mathrm{is_{PCT}} + 0.847 * \mathrm{nb_{claims}} + 0.456 * \mathrm{nat_{phase}} - 0.419 * \mathrm{reg_{phase}} + 0.335 * \mathrm{appln_{auth}} - 0.237 * \mathrm{duration} + 0.189 * \mathrm{nb_{local_{inventor}}} + 0.188 * \mathrm{nb_{local_{applicant}}} + 0.107 * \mathrm{family_{size}} + 0.099 * \mathrm{ratio_{granted}} - 0.096 * \mathrm{nb_{applicants}} + 0.087 * \mathrm{nb_{foreign_{inventor}}} - 0.071 * \mathrm{ctry_{first_{applicant}}} - 0.070 * \mathrm{nb_{inventors}} + 0.027 * \mathrm{nb_{applications}} + 0.027 * \mathrm{nb_{foreign_{applicant}}} - 0.022 * \mathrm{nb_{abstr_{char}}} + 0.005 * \mathrm{nb_{relevant_{patent}}} + 0.005 * \mathrm{nb_{title_{char}}} + 0.004 * \mathrm{tech_{field}} + 0.004 * \mathrm{nace_{code}} + 0.001 * \mathrm{backward_{citation}} - 1.270$$

Formula 2

Discussion and conclusion

This study introduces a novel algorithm for patent grant prediction based on the Kolmogorov-Arnold Network (EN-KAN), which enhances interpretability while maintaining superior performance. Unlike traditional multilayer perceptions, the proposed model leverages the Kolmogorov-Arnold theorem to overcome the limitations of conventional methods that rely on linear transformations combined with activation functions. By allowing the use of nonlinear functions, this approach provides a more detailed analysis of the nonlinear impacts of input variables on outputs, offering intuitive insights into decision-making processes. To validate the proposed model, we collected patent datasets from Electronic Communication fields and extracted potential influencing factors at different levels. To further improve the predictive performance, ensemble learning strategies were employed to enhance the model's generalization ability. The final trained model consistently outperformed traditional machine learning algorithms across multiple datasets, achieving performance levels comparable to neural networks. More importantly, the model provides feature importance rankings and directly generates equations, offering precise explanations for influential factors.

The findings reveal that the factors influencing patent grant exhibit significant consistency across fields, with examination-level and patent-level factors playing pivotal roles. Among examination- level factors, the submission of a PCT application shows a strong positive correlation with patent grants. This relationship is closely tied to the international, national, and regional phases, each of which serves distinct purposes in the patenting process. The international phase primarily focuses on patentability searches, providing applicants with more time to determine target markets. In contrast, the national and regional phases involve substantive reviews to secure patent protection in individual jurisdictions or regional organizations. Patent-level factors also significantly influence granting outcomes, with backward citation and the number of claims standing out as critical variables. Backward citation, which reflects the foundational knowledge underlying the innovation, is positively associated with patent grants, corroborating prior studies that link it to patent value (Junbyoung Oh & Wonchang Hur, 2018). The number of claims, often considered an indicator of patent scope (Novelli, 2015), displays an unexpected positive correlation with patent granting probabilities. This finding challenges the conventional view that more claims result in stricter examination processes and lower grant rates (Marco et al., 2019). Instead, the study aligns with recent research suggesting that the number of claims represents not only the scope but also the comprehensiveness and innovativeness of a patent, thereby highlighting its potential value (Kuhn & Thompson, 2019; Yao & Ni, 2023).

This study introduces the EN-KAN model, which combines interpretability with

high predictive performance. By leveraging the Kolmogorov-Arnold theorem instead of traditional multilayer neural network methods, the model not only identifies the key factors influencing patent granting but also provides mathematical formulas with coefficients. This approach addresses the "black box" problem inherent in neural network algorithms, further enhancing the interpretability of the predictive model. From a practical application perspective, these findings can assist innovative entities in optimizing their patent application strategies. Innovators in different fields can tailor their patent documentation based on their specific key factors, refine their patent portfolios, and significantly improve the likelihood of granting. For examination authorities, understanding the critical factors influencing patent granting enables a more focused review process, enhancing examination efficiency and refining patent review rules. Lastly, these conclusions can also guide research and market strategies. Considering patent grant factors during the research and development phase can facilitate the creation of technologies that are not only more patentable but also have higher market potential.

In summary, this study proposes EN-KAN as a robust tool for patent grant prediction, yet two limitations should be noted. First, the dataset used in this study is limited to a single technological domain and includes only patents filed in 2017, which may raise concerns regarding the generalizability of the findings. Future research could expand the scope to include multiple domains and application years to enable comparative analysis and enhance the robustness of the results. Additionally, despite efforts to include all relevant influencing factors, certain features, such as patent filing strategies, could not be incorporated due to data limitations. Future research could address this by exploring additional data sources to include a broader range of influencing factors.

Reference

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems, 34*, 4699–4711.

Alcácer, J., Gittelman, M., & Sampat, B. (2009). Applicant and examiner citations in U.S. patents: An overview and analysis. *Research Policy, 38*(2), 415–427.

Allen, G. I., Gan, L., & Zheng, L. (2024). Interpretable Machine Learning for Discovery: Statistical Challenges and Opportunities. In *Annual Review of Statistics and Its Application*, 11, 97–12.

Bekkers, R., Martinelli, A., & Tamagni, F. (2020). The impact of including standards-related documentation in patent prior art: Evidence from an EPO policy change. *Research Policy, 49*(7), 104007.

Costa, V. G., & Pedreira, C. E. (2023). Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, *56*(5), 4765–4800.

Covert, I. C., Lundberg, S., & Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *J. Mach. Learn. Res.*, *22*(1), Article 209.

Crawshaw, M., Liu, M., Orabona, F., Zhang, W., & Zhuang, Z. (2022). Robustness to unbounded smoothness of generalized signsgd. *Advances in Neural Information Processing Systems*, *35*, 9955–9968.

Desai, P. (2019). Biased Regulators: Evidence from Patent Examiners. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3485965

Dong, Y., Su, H., Zhu, J., & Bao, F. (2017). Towards interpretable deep neural networks by leveraging adversarial examples. *arXiv Preprint arXiv:1708.05493*.

Drivas, K., & Kaplanis, I. (2020). The role of international collaborations in securing the patent grant. *Journal of Informetrics*, *14*(4), 101093.

Frakes, M. D., & Wasserman, M. F. (2021). Knowledge spillovers, peer effects, and telecommuting: Evidence from the US Patent Office. *Journal of Public Economics*, *198*, 104425.

Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. *Ieee Access*, *9*, 144352–144360.

Gans, J. S., Hsu, D. H., & Stern, S. (2008). The Impact of Uncertain Intellectual Property Rights on the Market for Ideas: Evidence from Patent Grant Delays. *Management Science*, *54*(5), 982–997.

Guellec, D., & de la Potterie, B. van P. (2000). Applications, grants and the value of patent. *Economics Letters*, *69*(1), 109–114.

Harhoff, D., & Wagner, S. (2009). The duration of patent examination at the European Patent Office. *Management Science*, *55*(12), 1969–1984.

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., & Peste, A. (2021). Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, *22*(241), 1–124.

J. Dai, K. He, & J. Sun. (2015). Convolutional feature masking for joint object and stuff segmentation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3992–4000.

Junbyoung Oh & Wonchang Hur. (2018). A Man is Known by the Company He Keeps? : A Structural Relationship Between Backward Citation and Forward Citation of Patents. *Research Policy*, *50*(1), 104117.

Kuhn, J., & Thompson, N. (2019). How to Measure and Draw Causal Inferences with Patent Scope. *International Journal of the Economics of Business*, *26*, 5–38.

Lee, C., Kwon, O., Kim, M., & Kwon, D. (2018). Early identification of emerging technologies: A machine learning approach using multiple patent indicators. *Technological Forecasting and Social Change*, *127*, 291–303.

Lee, P.-C. (2021). Investigating Long-Term Technological Competitiveness: Originality, Generality, and Longevity. *IEEE Transactions on Engineering Management*, *71*, 20–42.

Lee, Y.-G., & Lee, J.-H. (2010). Different characteristics between auctioned and non-auctioned patents. *Scientometrics*, *82*(1), 135–148.

Lemley, M. A., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. *Review of Economics and Statistics*, *94*(3), 817–827.

Liegsalz, J., & Wagner, S. (2013). Patent examination at the state intellectual property office in China. *Research Policy*, *42*(2), 552–563.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, *16*(3), 31–57.

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., ... & Tegmark, M. (2024). Kan: Kolmogorov-arnold networks. arXiv preprint arXiv:2404.19756.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

Marco, A. C., Sarnoff, J. D., & Charles, A. W. (2019). Patent claims and patent scope. *Research Policy*, *48*(9), 103790.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv Preprint arXiv:1802.03426*.

Millar, C. C., Groth, O., & Mahon, J. F. (2018). Management innovation in a VUCA world: Challenges and recommendations. *California Management Review*, *61*(1), 5–14.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Niosi, J. (2006). Introduction to the Symposium: Universities as a Source of Commercial Technology. *The Journal of Technology Transfer*, *31*(4), 399–402. https://doi.org/10.1007/s10961-006-0001-0

Nordhaus, W. D. (1969). *Invention growth, and welfare; a theoretical treatment of technological change*. M.I.T. Press Cambridge, Mass.

Novelli, E. (2015). An examination of the antecedents and implications of patent scope. *Research Policy*, *44*(2), 493–507.

Raiteri, E. (2018). A time to nourish? Evaluating the impact of public procurement on technological generality through patent data. *Research Policy*, *47*(5), 936–952.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rong, S., & Bao-Wen, Z. (2018). The research of regression model in machine learning field. *MATEC Web of Conferences*, *176*, 01033.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sampat, B., & Williams, H. L. (2019). How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome. *American Economic Review*, *109*(1), 203–236. https://doi.org/10.1257/aer.20151398

Shen, M., & Zingg, R. (n.d.). *Patent Examiners and the Citation Bias in Innovation*.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(11).

Webster, E., Jensen, P. H., & Palangkaraya, A. (2014). Patent examination outcomes and the national treatment principle. *The RAND Journal of Economics*, *45*(2), 449–469.

Yang, D. (2008). Pendency and grant ratios of invention patents: A comparative study of the US and China. *Research Policy*, *37*(6–7), 1035–1046.

Yang, G., Lu, G., Xu, S., Chen, L., & Wen, Y. (2023). Which type of dynamic indicators should be preferred to predict patent commercial potential? *Technological Forecasting and Social Change*, *193*, 122637.

Yao, L., & Ni, H. (2023). Prediction of patent grant and interpreting the key determinants: An application of interpretable machine learning approach. *Scientometrics*, *128*(9), 4933–4969.

Zhou, Z. H., & Liu, S. (2021). *Machine Learning*. Springer Nature Singapore. https://books.google.ru/books?id=ctM-EAAAQBAJ

https://doi.org/10.51408/issi2025_067

Boost Formalism- A New Framework to Assess the Impact of Collaborations at Institutional Level

Prashasti Singh[1], Vivek Kumar Singh[2], Abhirup Nandy[3], Hiran H Lathabai[4]

[1]*prashasti.singh8@gmail.com*
Department of Computer Science, Shri Ram College of Commerce, University of Delhi, Delhi-110007 (India)

[2]*vivekks12@gmail.com*
Department of Computer Science, University of Delhi, Delhi-110007 (India)
NITI Aayog, Government of India, Delhi-10007 (India)

[3]*abhirupnandy.online@gmail.com*
Institute of Informatics and Communication, University of Delhi, Delhi-110067 (India)

[4]*hiranhl007@gmail.com*
Amrita CREATE, Amrita Vishwa Vidyapeetham, Amritapuri-690525, Kerala (India)

Abstract

Research collaboration at the international level has increased manifold during the last two decades. In addition to mutual benefits in the form of infrastructure sharing and knowledge flows, technology development and transfer, complementary and common solutions for shared problems, etc., research collaboration has also been associated with higher research productivity and impact. There are several previous studies that tried to measure and analyze international research collaboration for different countries and regions, and in the process developed different indicators and formalisms. However, there is no well-defined indicator to quantify the possible impact of international research collaboration on research output and citations of an institution. Recently, a set of boost indicators was introduced to reflect the effect of collaboration on productivity, impact, etc., of countries. This paper explores the possibility of adopting the boost formalism at an institutional level. The formalism is deployed and evaluated on research output data of 1000 Indian institutions. Different boost indicators are computed and validated through correlation studies. Results indicate that the proposed boost formalism can act as a suitable measure for assessing the possible impact of international research collaboration on research output and citations of institutions.

Introduction

Research collaboration is often defined as a group of researchers working together to solve complex scientific problems (Katz & Martin, 1997). It has also been defined as a social phenomenon where researchers pool their knowledge, experience, skills, and technology, intending to produce new scientific knowledge (Bozeman & Boardman, 2014). Research collaboration provides researchers with numerous mutual advantages in the form of knowledge transfer and training of researchers, resource sharing, access to complex and costly equipment, infrastructure, expansion and diversification of research network, funding etc. (Katz & Martin, 1997, Beaver, 2001; Birnholtz, 2007; D'Ippolito & Ruling, 2019). With the ICT revolution, the distances to interactions and collaborations have decreased, and as a result, the

research collaboration is now transcending institutional and geographical boundaries. Several studies have analyzed collaboration at the international level and have observed that it has risen linearly during the last two 2-3 decades, as measured in terms of the number of internationally co-authored papers published (Glanzel, 2001; Persson, Glänzel & Danell, 2004; Lee & Bozeman, 2005; Wagner & Leydesdorff, 2005; Leydesdorff & Wagner, 2008; Mattsson et al., 2008; Adams, 2012). Considering the benefits of International Research Collaboration (IRC), policymakers of different countries see it as a valuable tool and are designing various programs to foster such collaboration (Katz & Martin, 1997; Wagner et al., 2001; Boekholt et al., 2009).

Some studies have postulated that scientific collaboration is strongly associated with research productivity and economic growth, along with a significant impact on citation (Glänzel, 2001; Abramo, DÁngelo, & Solazzi, 2011; Abramo, DÁngelo & Murgia, 2017; Inglesi-Lotz & Pouris, 2013; Ntuli et al., 2015). Many previous studies have focused their attention on measuring and characterising IRC trends, patterns, and impacts in different countries. Various indicators to measure the association strength in terms of propensity, intensity, and affinity in international collaboration have been proposed. Initially, the key focus of the analysis of IRC was the size of the country and related geographical, socioeconomic, and historical factors (Price, 1969; Frame & Carpenter, 1979) shaping research collaboration. As the research work on IRC grew, some indicators like the 'cooperation index' (Schubert & Braun, 1990) and 'exclusive strategy' (Luukkonen, Persson & Sivertsen, 1993) were introduced. The weighted affinity index was introduced thereafter (Leclerc & Gagné, 1994) to weigh the measured links between two countries based on the observed/expected ratio. For calculating absolute strength between pairs of countries, the Salton measure was proposed (Schubert & Braun, 1990; Glanzel, 2001). In addition to the affinity index, similarity measures such as cosine similarity (van Eck & Waltman 2009), inclusion index (van Eck & Waltman 2009; Luukkonen, 1993), Jaccard similarity (van Eck & Waltman 2009; Luukkonen, 1993), and multilateral similarity (Goodman's quasi-independence) (Luukkonen, 1993) were applied to bibliometric data. Three major algorithms have been proposed to define the Probability Affinity Index (PAI), namely non-overlapping (Leclerc and Gagné, 1994), overlapping (Zitt et al, 2000), and self-exclusive methods (Luukkonen, Persson & Sivertse, 1992; Schubert & Glänzel, 2006). The partnership probability index was developed by Yamashita & Okubo (2006) and was applied in combination with PAI as the Salton-Ochiai index on inter-sectoral organizational collaboration. Recently, some variants of the relative intensity of collaboration were studied by Fuchs, Sivertsen & Rousseau (2021).

Though many of the previous studies proposed indices to measure and characterize international research collaboration, there has not been a development towards a suitable indicator to measure the impact of international research collaboration on the research output and citations of an institution, a country, or any other actor in the scientific research landscape. There lies the research gap that this study attempts to bridge. The study proposes a Boost formalism consisting of different boost indicators that can be used to measure what effect or impact the international research

collaboration may have on the productivity (research output is the proxy taken) or impact (citations are the proxy taken) of an institution. The formalism is described in detail, and thereafter its applicability in an institutional context is demonstrated on research publication data of 1000 Indian institutions. The suitability and relevance of boost indicators are evaluated. Finally, the usefulness, applicability, and further extension possibilities of formalism are discussed.

Related work

The investigation of international research collaboration (IRC) through co-authorship patterns began with efforts to characterize the interaction between the scientific output of a nation and its large-scale determinants. Price's (1969) contribution was a path-breaking effort in this regard, where he analyzed the correlations between a nation's scientific activities and socioeconomic determinants such as economic scale and technological capability. This initial effort brought to the forefront the role of national resources in shaping the dynamics of scientific collaboration. Frame and Carpenter (1979) took these findings further by examining the 1973 Science Citation Index (SCI) data, which included over 100 subfields categorized into nine scientific fields across 167 countries. The study found a positive correlation between a nation's scientific capability, measured by publication output, and internationally co-authored publications, indicating that larger scientific communities engage more in global collaborations. These early studies formed the foundation for systematic methodologies in IRC measurement, with the role of national scale in shaping collaboration behavior being a central theme.

To measure collaboration strength and trends, researchers have come up with various indicators to quantify IRC. Schubert and Braun (1990) came up with the cooperative index, a percentage difference between actual international co-authorships and expected values, adjusted for country size. The index made it possible to compare collaboration tendencies between countries, taking into consideration differences in scientific output. They also used Salton's measure (Salton & McGill, 1983), which measures the relative intensity of co-authorship relationships between countries. The measure was used by Glänzel and Schubert (2001) to analyze collaboration between 36 countries. Though useful for symmetric collaboration patterns, Salton's measure is difficult to use to capture asymmetric relationships, where a country dominates the partnership, and thus is of limited use in various collaboration scenarios. Luukkonen, Persson, and Sivertsen (1992) responded to size-dependency with the Probabilistic Affinity Index (PAI), which attempts to quantify collaboration strength regardless of country size. PAI cross-checks actual co-authorships against expected ones, and values above 1 represent stronger-than-expected collaboration. PAI, however, overestimates the importance of countries with skewed collaboration distributions and those with dominant partners. To counteract this, Schubert and Glänzel (2006) created the preference index of co-authorship, which is an enhancement on PAI in the sense that it accounts for specific country collaboration preferences and removes size effects. This index generates a more advanced measure of bilateral scientific connections, reflecting the country's affinity more precisely. Luukkonen et al. (1993) also suggested other measures of collaboration intensity, such as bilateral similarity

measures (e.g., Jaccard, Salton), multilateral similarity by Goodman's quasi-independence model, and multidimensional scaling for graphical representation of IRC networks. Such methods, though pioneering, remain size-dependent, overestimating the contribution of large countries compared to small ones, which makes equitable comparisons difficult.

Later studies developed new indicators to overcome earlier limitations. Leclerc and Gagné (1994) developed the proximity index (PRI), a quantifier of the strength of collaboration against the number of co-authored outputs. The PRI is aimed at symmetric relationships between nations, with greater values signifying stronger collaborative relations; however, its focus on symmetry limits its use. Zitt, Bassecoulard, and Okubo (2000) developed a publication-level probabilistic affinity index, in contrast to the co-authorship-level PAI, to measure the strength of collaboration between five major scientific nations: France, Germany, Japan, the UK, and the USA. Their approach overcame the impact of self-co-authorship through iterative margin recalibrations, thus ensuring a fair assessment of international relations. Yamashita and Okubo (2006) examined inter-sectoral collaboration between France and Japan through the combination of PAI with Salton's measure, a modification of the Ochiai coefficient (Ochiai, 1957; Zhou & Leydesdorff, 2016). They also developed the Probabilistic Partnership Index (PPI), measuring the infrequency of observed partnership links against predicted distributions. The PPI complements the PAI by identifying the statistical significance of partnerships, thus introducing a new dimension to collaboration processes.

Recent advances have focused on improving IRC measures to address contemporary challenges. Fuchs, Sivertsen, and Rousseau (2021) introduced the Relative Intensity of Collaboration (RIC), an improvement over earlier asymmetric indices, such as Luukkonen's PAI, which failed to capture relative increases in co-authored papers (Rousseau, 2021). RIC provides a robust measure of collaboration intensity by considering total collaboration volumes and pairwise interactions, thus improving its performance in asymmetric cases. Chinchilla-Rodriguez et al. (2021) explored differences in the use of PAI, such as differences in the handling of co-authorship matrix diagonals (e.g., setting to zero, as in Luukkonen et al., 1992; Leclerc & Gagné, 1994; Schubert & Glänzel, 2006; Fuchs et al., 2021) and normalization methods (Zitt et al., 2000; Yamashita & Okubo, 2006). These differences show the complexity of standardizing IRC measures across different research environments.

Counting methods have also been included in IRC analysis, providing authorship credit in collaborative research. Full counting provides equal credit to all authors, while fractional counting provides proportionate credit (Frandsen & Nicolaisen, 2010; Harsanyi, 1993; Lindsey, 1980; Waltman, 2016). Gauffriau (2017) outlined these approaches, highlighting their strengths and weaknesses in bibliometric studies. Most PAI-based analyses employed full counting, except for Leclerc and Gagné (1994) and Zitt et al. (2000), which explored fractional alternatives. Braun, Glänzel, and Schubert (1991) and Okubo, Miquel, Frigoletto, and Doré (1992) also dealt with the implications of counting methods for fair collaboration assessment.

As much as IRC indicators are prevalent across the world, there is an urgent gap: there is no measure among the current ones that reflects the impact of IRC on institutional productivity (publication output) or influence (citations). While country-level evidence has been useful, evidence at the institutional level is required to know how collaborations define research landscapes. This paper fills this gap by introducing a boost formalism—a collection of straightforward indicators to approximate the impact of IRC on institutional citations and publications. Using publication data from 1,000 Indian institutions, this framework offers a new approach to guide institutional strategies and policymaking, complementing traditional bibliometric measures.

Boost formalism: A discussion

Dua et al. (2023) introduced a set of indicators, viz. the boost indicators, to reflect the effect of collaborations on productivity, impact, etc., of countries. The idea of a boost in productivity and citation provides a way to quantify the impact of collaboration on productivity, citations, and altmetrics for different countries. The boost measures can be extended to the institutional context as follows:

Productivity boost **(β_P):** It can be defined as the ratio of the total number of publications (TP) to the total number of indigenous publications (TIP) of an institution, expressed in percentage. It can be expressed as follows,

$$\beta_P = \left[\frac{TP}{TIP} - 1\right] \times 100\ \%$$

The expression suggests that if an institution does not engage in collaboration, then $\beta_P = 0\ \%$. The value of β_P is directly proportional to the boost in productivity due to collaborations. A higher value of β_P indicates a higher reliance of the institution on international research collaboration. The ideal value of β_P is difficult to determine. As per the rule of thumb, if $\beta_P > 50\ \%$, then it indicates that the institution is more dependent on international collaboration than the indigenous ecosystem. On the other hand, if $\beta_P > 100\ \%$, it indicates that the institution is highly dependent on collaboration. If an institution has an infinite β_P (TIP=0 and a TP value of 1 or above), it signifies absolute dependence on collaboration.

Citations boost **(β_C):** It is defined as the ratio of total citations (TC) to the total citations received by indigenous publications (TIC) of an institution.

$$\beta_C = \left[\frac{TC}{TIC} - 1\right] \times 100\ \%$$

As per the rule of thumb, if $\beta_C > 50\ \%$, then it indicates the institution is more reliant/dependent on international collaborations for citation or impact than the indigenous scholarly system. On the other hand, if $\beta_C > 100\ \%$, it indicates that the institution is highly dependent. In other words, this indicates that the indigenous scholarly ecosystem is drawing very low relative impact and reach. Therefore, the

institution should choose some impactful platforms or sources to disseminate its scientific research and improve the visibility of the indigenous scholarly research outputs.

Boost ratio of impact per unit boost in productivity **(γ_c):** It is the net boost of citation per unit boost of productivity due to international research collaborations.

$$\gamma_c = \frac{\beta_C}{\beta_P}$$

If the value of $\gamma_c < 1$, international research collaborations are less rewarding and if $\gamma_c > 1$, such collaborations are rewarding. The benefit of research collaboration depends on the value of γ_c. This means the higher the value of γ_c, the greater the benefit of collaboration.

Citedness boost **(β_{rc}):** It is the ratio of total citedness (total cited ratio) to the citedness ratio of the indigenous publications.

$$\beta_{rc} = \left[\frac{r_T}{r_{TI}} - 1\right] \times 100\ \%$$

where

$$r_T = \frac{total\ number\ of\ cited\ publications}{total\ number\ of\ publications} = \frac{TP_{cited}}{TP}$$

&

$$r_{TI} = \frac{total\ number\ of\ cited\ indigenous\ publications}{total\ number\ of\ indigenous\ publications} = \frac{TIP_{cited}}{TIP}$$

Citedness boost value greater than but close to 1 indicates that indigenous publications also have considerably good citedness. β_{rc} and β_c can be used together to determine whether an institution's indigenous works are making enough impact. β_{rc} value closer to 1 (like < 1 %), but considerably high β_c (like > 50 %) can indicate that despite the potential of indigenous works to gain citations, a considerable amount of work is remaining under-cited or not getting enough citations.

Boost ratio of impact per unit boost in citedness **(δ_c):** It is the net boost of impact per unit boost of citedness due to international collaborations.

$$\delta_c = \frac{\beta_C}{\beta_{rc}}$$

The effectiveness of collaboration depends on the value of δ_c. The higher the value of δ_c, the higher the effectiveness of foreign collaboration. If the value of δ_c is very high with $\beta_{rc} < 1\ \%$, it indicates that the majority of collaboration is of good

quality and rewarding as well. On the other hand, a high value of δ_c with $\beta_{rc} > 1$ %, indicates that there are some less rewarding collaborations. The reason for this could be that the collaboration can be a new tie or maybe the collaboration was formed long back but working on obsolete themes. Therefore, such collaboration should be reviewed to strengthen the collaboration by working on trending themes, to stop weaker ties and search for new ties or to minimize emphasis on such collaboration.

Demonstration of the Formalism

Data

In order to demonstrate the formalism of Boost in productivity and citations, research publication data for a large set of 1,000 Indian institutions collected from the Dimensions for an earlier work (Singh *et al.*, 2022) was used. The top 1000 Indian Institutions were selected on the basis of the total research output of those institutions during 2010-2019. The data comprised all document types and corresponded to the time period 2010 to 2019. The metadata fields that were accessed included the year of publication, DOI, citations, author(s) country affiliation, etc. The query formulated was as follows:

Search Query
search publications where year in [2010:2019] and research_orgs.id="{GRIDID}" and type in ["article"] return publications [research_org_countries+type+authors+year+abstract+open_access_categories_v2+research_orgs+authors_count+concepts_scores+field_citation_ratio+publisher+times_cited+altmetric_id+category_for+doi+title+category_sdg+journal+reference_ids+id+altmetric+issn+funder_countries+funders+relative_citation_ratio+supporting_grant_ids]

In the search query above, "GRIDID" corresponds to a unique ID assigned to each institution and these IDs for the top 1000 Indian Institutions were taken from the database. This was then passed one by one in the search query post which data for each of the Institutions was downloaded and processed.

Methodology

Post data download, different scientometric measures were computed by processing the appropriate metadata fields in the processed data. *Firstly*, the values of TP (total papers) and TC (total citations) were computed. TP was obtained from the total data count for each institution, while TC was obtained by summing up the values under the "times_cited" field for each institution. *Secondly*, in order to get the count of ICP (internationally collaborated papers) the "research_org_countries" field was investigated. This field contained the names of countries that collaborated to publish a record. Thus, for each institution, ICP comprised the total number of records that had more than country (India) listed in this field; while the records that had only one country (India) listed in this field comprised the share of TIP (total number of indigenous publications). Similarly, TIC (total number of indigenous citations) was

obtained by summing up the values under the "times_cited" field that corresponded to only one country (India) listed in the "research_org_countries" field. *Thirdly*, for each institution, the computed values of productivity and citations were then used to compute the different boost indicators mentioned above. *Finally*, to better realise the nature of the different computed boost indicators, their values were correlated with the NIRF (National Institution Ranking Framework) ranks of each Indian Institution. A brief overview on NIRF is provided in the *Evaluation* section of this paper.

Results

The different boost indicators were computed for all the 1,000 institutions considered. The values for a set of 50 such institutions having high research output are presented in **Table 1**. The file containing the complete list of the 1,000 institutions considered, along with their relevant values and computations would be provided on request.

Table 1. Different Productivity Indicators of selected 50 Institutions.

S. No.	Institution Name	Acronym	TP	TIP	ICP	ICP %	TC	TIC	βp	βc	Ɣc	βrc
1	Anna University, Chennai	AU Chennai	29698	25995	3703	12.47	316029	243145	14.245	29.976	2.104	1.918
2	All India Institute of Medical Sciences, Delhi	AIIMS Delhi	20545	17869	2676	13.03	225624	124171	14.976	81.704	5.456	2.498
3	Indian Institute of Science Bangalore	IISC	20257	15004	5253	25.93	308491	198107	35.011	55.719	1.591	1.516
4	Indian Institute of Technology Kharagpur	IIT KGP	18329	14621	3708	20.23	274172	200985	25.361	36.414	1.436	1.003
5	Indian Institute of Technology Bombay	IITB	17384	12861	4523	26.02	235472	149501	35.168	57.505	1.635	1.98
6	Indian Institute of Technology Madras	IITM	16650	12877	3773	22.66	207338	145541	29.3	42.46	1.449	1.415
7	Indian Institute of Technology Delhi	IITD	15402	12211	3191	20.72	232485	165499	26.132	40.475	1.549	1.114
8	University of Delhi	DU	15134	11288	3846	25.41	235350	123865	34.072	90.005	2.642	4.684
9	Bhabha Atomic Research Centre	BARC	13752	10443	3309	24.06	207521	121162	31.686	71.276	2.249	1.132
10	Post Graduate Institute of Medical Education and	PGIMER Chandigarh	13712	12224	1488	10.85	142646	91401	12.173	56.066	4.606	2.009

	Research, Chandigarh											
11	Vellore Institute of Technology University	VITU	12526	10072	2454	19.59	150495	107739	24.365	39.685	1.629	2.141
12	Jadavpur University	JU	12502	10216	2286	18.29	167053	120836	22.377	38.248	1.709	1.423
13	Indian Institute of Technology Roorkee	IITR	12470	10089	2381	19.09	207482	144100	23.6	43.985	1.864	0.756
14	Indian Institute of Technology Kanpur	IITK	11583	8776	2807	24.23	154648	102439	31.985	50.966	1.593	1.785
15	Indian Institute of Technology Guwahati	IITG	10364	8432	1932	18.64	146145	103298	22.913	41.479	1.81	1.252
16	Banaras Hindu University	BHU	10214	7959	2255	22.08	176674	117377	28.333	50.518	1.783	1.716
17	University of Calcutta	CU	9703	7850	1853	19.1	113064	81394	23.605	38.91	1.648	1.791
18	University of Pune	SPPU	9510	8046	1464	15.39	99732	72156	18.195	38.217	2.1	2.311
19	Visvesvaraya Technological University, Belgaum	VTU Belgaum	8959	7937	1022	11.41	64177	50249	12.876	27.718	2.153	2.058
20	Panjab University	PU	8469	5616	2853	33.69	171093	78320	50.801	118.454	2.332	3.538
21	Manipal Academy of Higher Education, Manipal	MAHE	8307	6575	1732	20.85	74606	48989	26.342	52.291	1.985	3.054
22	Aligarh Muslim University	AMU	8025	5744	2281	28.42	123656	73717	39.711	67.744	1.706	2.388
23	Maulana Azad National Institute of Technology, Bhopal	MANIT Bhopal	7866	6859	1007	12.8	107564	85646	14.681	25.591	1.743	0.824
24	University of Madras	UNOM	7017	5573	1444	20.58	96803	64557	25.911	49.95	1.928	2.616
25	University of Hyderabad	HCU	6651	5288	1363	20.49	91303	62711	25.775	45.593	1.769	2.999
26	Indian Institute of Chemical Technology, Hyderabad	IICT	6519	5485	1034	15.86	103422	75896	18.851	36.268	1.924	4.643
27	Jawaharlal Nehru University	JNU	6363	5068	1295	20.35	91933	54243	25.552	69.484	2.719	4.308
28	Amity University, Noida	AUUP	6325	5036	1289	20.38	57284	37570	25.596	52.473	2.05	2.663

29	Indian Institute of Technology (ISM) Dhanbad	ISM	6322	5552	770	12.18	78416	64986	13.869	20.666	1.49	0.977
30	Bharathiar University, Coimbatore	BU Coimbatore	6194	4197	1997	32.24	88737	46165	47.582	92.217	1.938	4.166
31	Tata Institute of Fundamental Research	TIFR	6152	2564	3588	58.32	143789	26907	139.938	434.393	3.104	6.908
32	University of Kerala	UK	5834	5058	776	13.3	54766	42393	15.342	29.186	1.902	1.993
33	Annamalai University	AU Tamil Nadu	5376	4447	929	17.28	80872	63250	20.89	27.861	1.334	1.443
34	Christian Medical College & Hospital, Vellore	CMCH Vellore	5334	4067	1267	23.75	56797	24848	31.153	128.578	4.127	4.76
35	Pondicherry University	Pondicherry University	5064	4257	807	15.94	60962	44344	18.957	37.475	1.977	2.271
36	King George's Medical University, Lucknow	KGMU Lucknow	5050	4622	428	8.48	37116	29030	9.26	27.854	3.008	1.713
37	Thapar University, Patiala	TIET Patiala	4987	4245	742	14.88	74715	56080	17.479	33.229	1.901	0.975
38	Bharathidasan University	Bharathidasan University	4954	3577	1377	27.8	72566	45478	38.496	59.563	1.547	2.058
39	Jamia Milia Islamia	JMI	4923	3569	1354	27.5	77521	50081	37.938	54.791	1.444	3.066
40	National Institute of Technology Rourkela	NITR	4897	4363	534	10.9	60716	48926	12.239	24.098	1.969	0.448
41	Birla Institute of Technology and Science, Pilani	BITS Pilani	4774	3913	861	18.04	55937	39727	22.004	40.803	1.854	1.292
42	Indian Statistical Institute, Kolkata	ISI Kolkata	4751	3124	1627	34.25	52270	29198	52.081	79.019	1.517	2.104
43	Sanjay Gandhi Post Graduate Institute of Medical Sciences, Lucknow	SGPGI Lucknow	4652	4201	451	9.69	70416	33205	10.736	112.064	10.439	2.218
44	National Chemical Laboratory, Pune	NCL Pune	4598	3794	804	17.49	92440	65842	21.191	40.397	1.906	1.997
45	Indian Association for the Cultivation of	IACS Kolkata	4477	3501	976	21.8	83081	61666	27.878	34.727	1.246	1.131

	Science, Kolkata											
46	Indian Institute of Engineering Science and Technology, Shibpur	IIEST Shibpur	4438	3838	600	13.52	44685	36046	15.633	23.967	1.533	0.992
47	National Institute of Mental Health and Neurosciences, Bengaluru	NIMHANS	4416	3648	768	17.39	45522	28897	21.053	57.532	2.733	2.068
48	National Institute of Technology Tiruchirappalli	NIT-T	4353	3642	711	16.33	59008	46617	19.522	26.58	1.362	1.816
49	West Bengal University of Technology, Kolkata	MAKAUT WB	4351	3694	657	15.1	41509	32230	17.786	28.79	1.619	1.577
50	Amrita Vishwa Vidyapeetham University	AMRITA	4052	3479	573	14.14	41548	27121	16.47	53.195	3.23	1.816

Note: **TP**-> Total Publications, **TIP**-> Total Indigenous Publications, **TC**-> Total Citations, **TIC**-> Total Indigenous Citations, **ICP**-> Internationally Collaborated Publications (ICP=TP-TIP, ICP%=(ICP/TP)*100).

From **Table 1**, it can be observed that among the top 50 productive institutions, except for the Tata Institute for Fundamental Research, all institutions have βp values <50%, indicating a self-reliant research ecosystem. Also, 9 Institutions have βp values <15% which indicates that the institutions have achieved a much higher productivity boost through domestic publications without much need for collaborations which is also seen owing to the fact that their Internationally collaborated publications (ICP=TP-TIP) comprise a share of <15% of their Total Publications (TP). These institutions are, namely, AU Chennai, AIIMS Delhi, PGIMER Chandigarh, VTU Belgaum, MANIT Bhopal, ISM Dhanbad, KGMU Lucknow, NITR and SGPGI Lucknow. It is to be noted that among the 7 IITs appearing in the top 50 list, IIT Gandhinagar (Rank 15), IIT Roorkee (Rank 13), IIT Kharagpur (Rank 4), IIT Delhi (Rank 7) and IIT Madras (Rank 6) have βp values <30% while IIT Kanpur (Rank 14) and IIT Bombay (Rank 5) display βp values of approx 32% and 35% respectively while they rank much higher in terms of TP. Moreover, the minimum value of βp is observed for King George Medical University (KGMU Lucknow, 9.26%) while the maximum value of βp is observed for Tata Institute for Fundamental Research (TIFR, 139.938). However, in terms of ranking by TP, KGMU (Rank 36, TP 5050) ranks lower than TIFR (Rank 31, TP 6152). According to the interpretation of βp values, this indicates that TIFR, despite having published a greater number of research publications than KGMU Lucknow, is more

dependent on collaborative research than indigenous research, as the βp value for KGMU is >100%.

In terms of a Boost in productivity due to Citations, i.e. βc values, 28 Institutions have βc values <50%. A few of these are AU Chennai (~30%), IIT KGP (~36%), IIT Madras (42.5%), IIT Delhi (~40.48%), VITU (~40%), etc. Only 3 Institutions have βc values <25%, namely ISM (20.7%), IIEST Shibpur and NIT Rourkela (~24%). The maximum βc value is observed again for TIFR (434.39%), and the minimum is observed for ISM. It is to be noted that while TIFR (Rank 31, TP 6152) and ISM (Rank 29, TP 6322) differ marginally in terms of ranking due to TP, they lie on extreme ends of βc values. Thus, according to the interpretation of the βc values, the boost in citations achieved for TIFR is largely a result of its collaboration, while for ISM, it indicates a strong domestic research environment. As for the IITs appearing in the top 50 list, IIT Kanpur (50.97%) and IIT Bombay (57.5%) have βc values >50% while the other IITs like IIT Delhi (~40.48%), IIT Gandhinagar (~41.48%), IIT Madras (42.46%) and IIT Roorkee (~43.99%) have βc values <50%.

In terms of citedness boost i.e. βrc, 6 institutions (NITR, IITR, MANIT Bhopal, TIET Patiala, ISM and IIEST Shibpur), achieve values <1% which indicates impactful indigenous work by these institutions. These institutions also have βc values <50% which further supplements this finding. Among the IITs, it is seen that though IIT Roorkee (βrc=0.756, βc=43.98) has a lesser value of βrc than IIT KGP (βrc=1.003, βc=36.41) but has a higher value of βc than IIT KGP. On the other hand, IISC Bengaluru which ranks 3rd in terms of TP has both βrc=1.5% and βc=55.7% which indicates that both the boost in citations and the citedness boost are a result of collaborations. Here also, TIFR demonstrated the highest value of βrc i.e. 6.9%. Lastly, in terms of the Boost ratio of impact per unit boost in productivity (γ_c), almost all institutions have values > 1% which indicates that the international collaborations have been rewarding.

Cutoff values for βp (>50%, >100%), βc (>50%, >100%), and βrc (≈1%) were chosen using previous research patterns (Adams, 2012; Larivière et al., 2015). For βp >50% (TP = 1.5 × TIP, 33% of output) shows notable collaboration help, while >100% (TP = 2 × TIP) means heavy reliance. For βc >50% (TC = 1.5 × TIC) indicates collaboration boosts citations significantly. For βrc ≈1% means local and collaborative papers are cited similarly. **Table 1** shows KGMU's βp = 9.26% (self-reliant), TIFR's βp = 139.94% (TP = 2.4 × TIP), and IIT Roorkee's βrc = 0.756%. Figure 1's weak link ($R^2 = 0.0014$) supports these cutoffs.

To understand the relationship of the boost indicators with publication and citation counts, scatter plots of TP vs. βp, TC vs. βc, and TC vs. βrc are provided in **Figures 1**, **2** and **3**, respectively. **Figure 1** shows a very weak positive correlation ($R^2 = 0.0014$) between total publications (TP) and productivity boost (βp), suggesting that institutions with a high TP do not necessarily have a proportionally high βp. This implies that some institutions maintain strong indigenous publication ecosystems while others rely heavily on international collaborations. Notable institutions such as IISC, IITs, and AIIMS Delhi have a large TP but moderate βp, indicating a well-developed domestic research ecosystem. In contrast, institutions like ISI Kolkata, PU, and BU Coimbatore exhibit a high βp (>40%), signifying substantial reliance on

international collaborations. **Figure 2** examines how total citations (TC) relate to βc, which quantifies the citation boost received due to international collaboration. A low R^2 value (0.0115) in the trend line suggests a weak correlation, indicating that while international collaboration generally increases citations, the extent of this boost varies significantly across institutions. Institutions with high TC but moderate βc, such as IITs and IISc, indicate that their indigenous research is also widely cited. On the other hand, CMCH Vellore, SGPGI Lucknow, and BU Coimbatore, with high βc values (>90), heavily rely on internationally collaborated research for citations, implying that their domestic publications receive comparatively lower impact. **Figure 3** explores how citedness boost (βrc) varies with total citations (TC). A weak negative correlation ($R^2 = 0.0086$) indicates that institutions with high TC do not necessarily have a high βrc, signifying that indigenous research in certain institutions is already well cited. DU, IICT, JNU, and BU Coimbatore show higher βrc values (>4), meaning their internationally collaborated publications receive significantly more citations per paper than indigenous ones. In contrast, institutions like IISc, IITs, and BARC have moderate βrc, suggesting a relatively balanced impact between international and domestic publications. Finally, **Figure 4** shows the strongest correlation ($R^2 = 0.3162$) between βp and βc, indicating that institutions that achieve higher productivity boosts through international collaborations tend to also receive proportionally higher citation boosts. This moderate positive correlation indicates that the advantages of international collaboration typically appear simultaneously in both increased productivity and citation impact, although the degree of effect differs significantly among institutions. Notably, institutions such as CMCH Vellore, PU, and BU Coimbatore exhibit particularly strong performance in both metrics.

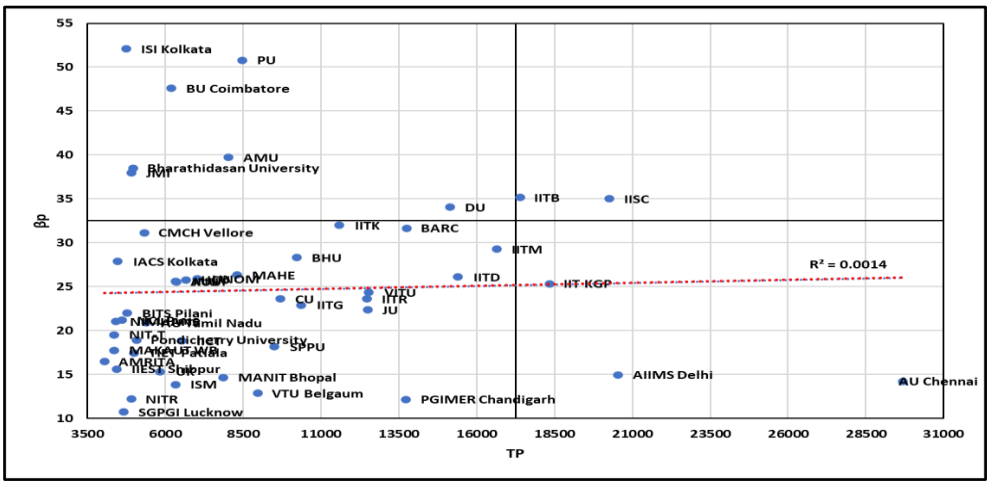


Figure 1. Boost in Productivity vs. Total Publications (TP) of selected 50 Institutions (excluding an outlier- TIFR).

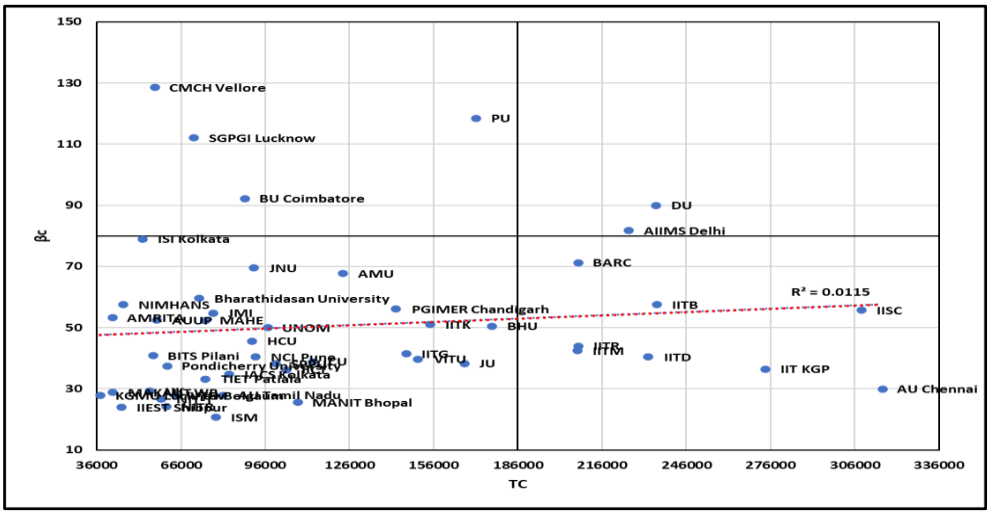


Figure 2. Boost in Citations vs. Total Citations (TC) of selected 50 Institutions (excluding an Outlier- TIFR).

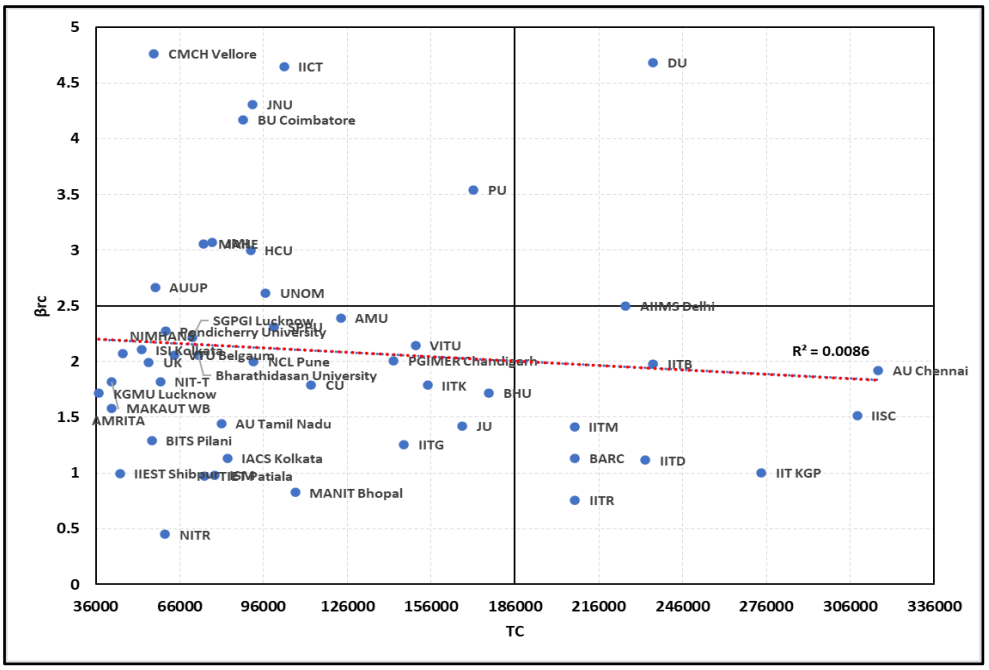


Figure 3. Boost in Citedness vs. Total Citations (TC) of selected 50 Institutions (excluding an Outlier- TIFR).

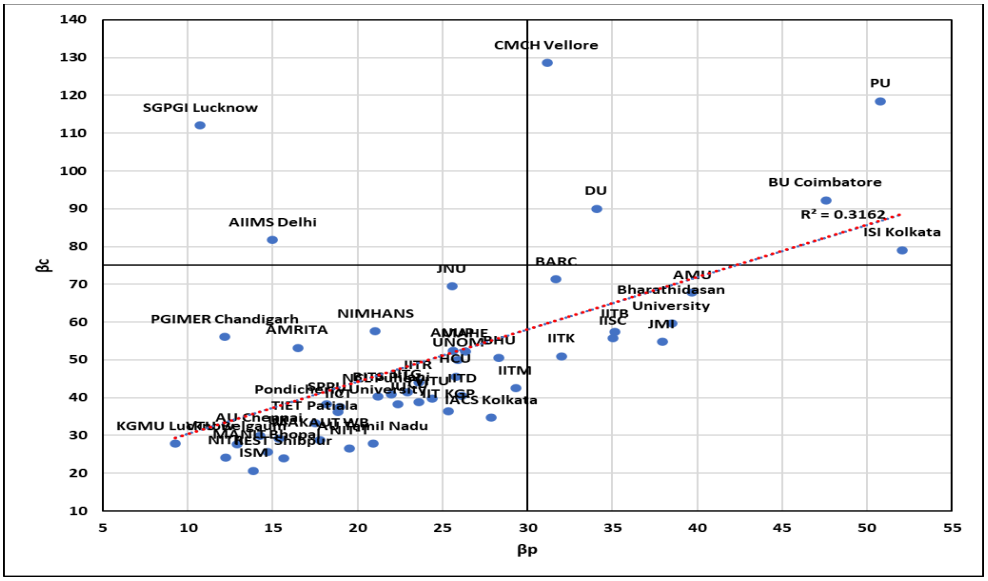


Figure 4. Boost in Citations vs. Boost in Productivity of selected 50 Institutions (excluding an outlier- TIFR).

Evaluation

In order to understand the nature and values of different boost indicators, the values of different boost indicators are correlated with a major national ranking of institutions in India, the NIRF. The National Institutional Ranking Framework (NIRF), established by India's Ministry of Education, ranks higher education institutions based on five weighted parameters: Teaching, Learning, and Resources (30%); Research and Professional Practices (30%); Graduation Outcomes (20%); Outreach and Inclusivity (10%); and Perception (10%). The research component, emphasising publication and citation metrics, aligns closely with the boost indicators' focus on productivity and impact, making NIRF a robust benchmark for validation.

The Spearman's Rank Correlation Coefficient (SRCC) was calculated to compare the rankings of higher education institutions by the NIRF with rankings by different productivity measures. The results from **Table 2** show a positive correlation with NIRF rankings and rankings by Total Publications (TP) and Total Citations (TC) at SRCC values of 0.67 and 0.68, respectively. The high correlation is due to the convergence of TP and TC with NIRF's high weighting of the research and professional practices criterion, which contributes 30% to the overall ranking methodology.

On the other hand, the correlations derived between NIRF rankings and the boost indicators—βp (0.27), βc (0.19), and βrc (0.03)—are significantly lower. These low correlation values indicate that these boost indicators might not be able to capture the complex nature of NIRF's ranking factors, which include Teaching, Learning, and Resources (30%), Graduation Outcomes (20%), Outreach and Inclusivity (10%), and Perception (10%).

This is explored further through **Figure 5**, which graphically verifies these findings. This is utilized to further support the argument that institutions with high publication and citation values rank better in NIRF's research-oriented evaluation. Conversely, the lower correlations of the boost indicators (βp, βc, and βrc) are evident from the less intense shading and the lower SRCC values, ranging from 0.03 to 0.27. This difference suggests that, while TP and TC are effective indicators of NIRF's focus on research productivity, the boost indicators might be measuring different aspects of institutional performance that are less aligned with NIRF's integrative approach.

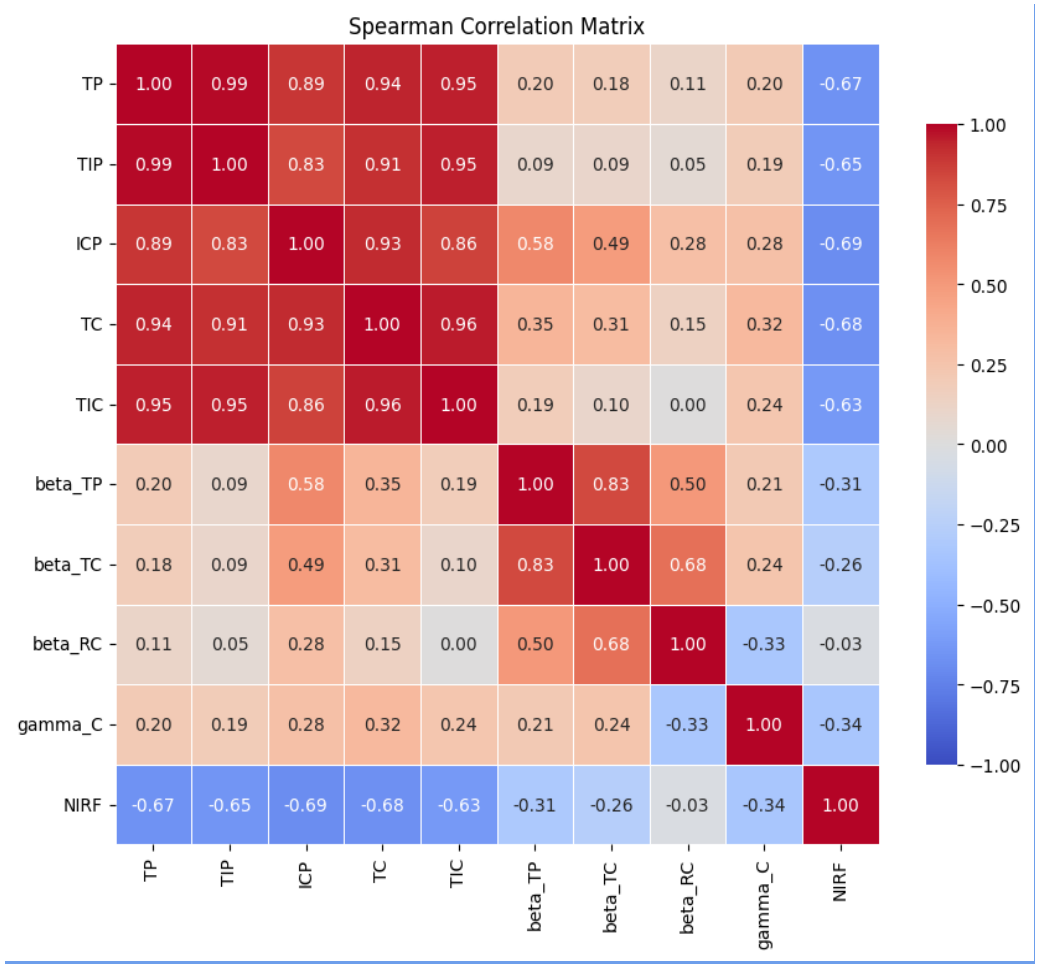


Figure 5. Spearman Correlation between all the parameters, boost parameters, and the NIRF rankings.

Table 2. Values of Spearman Rank Correlation Coefficients for different variables.

Variables	Value of SRCC
NIRF Ranking vs TP	0.67
NIRF Ranking vs TC	0.68
NIRF Ranking vs βp	0.27
NIRF Ranking vs βc	0.19
NIRF Ranking vs βrc	0.03

Discussion

This study analyses research publication data from 1,000 Indian institutions to assess the impact of research Collaboration using the Boost formalism. By adopting the boost indicators – βp (productivity boost), βc (citation boost), βrc (citedness boost), and ϒc – to an institutional level, this work offers a novel approach to quantifying collaboration effects beyond traditional bibliometric measures like Total Publications (TP) and Total Citations (TC).

The findings of the study reveal that collaboration influences institutions differently. While IITs and AIIMS Delhi maintain strong indigenous research ecosystems with moderate βp values, institutions like TIFR exhibit high βp and βc, suggesting greater dependence on collaborations. Weak correlations between TP and βp, as well as TC and βc, indicate that high publication volume does not always correspond to significant collaborative impact. However, a moderate positive correlation between βp and βc suggests that well-integrated collaborations enhance both productivity and citation impact.

The study has practical implications for institutional research profiling, academic planning, and policymaking. This is especially important because though (i) national scholarly ranking initiatives like NIRF provide a sense about their relative performance (ii) recently proposed indicators like x and x_d provides an idea about the scholarly research portfolio of institutions, these are not capable of providing an idea about the role and extent of influence collaborations have in determining the institutions' current stature. As a boost indicator, such as input, it can complement the information provided by NIRF and other useful indicators like x, x_d, and many others for institutions to plan their way forward and shape their research policy and formulate strategies. Institutions with high βc but moderate βp benefit from selective, high-impact partnerships, while those with high βp but low βrc may need to improve the visibility of their indigenous research. Policymakers can use these insights to allocate resources and design policies that foster meaningful collaboration. The profile of collaboration's impact on institutions highlights its dual role in enhancing productivity and global visibility. Institutions with strong international ties often gain access to cutting-edge knowledge, advanced

methodologies, and prestigious networks, contributing to their academic standing. Additionally, IRC enables researchers to tackle complex, multidisciplinary problems requiring diverse expertise, boosting institutional research output and reputation. Such collaborations also help institutions attract better funding, international faculty, and students, creating a virtuous cycle of growth and recognition in the global academic landscape.

Despite its contributions, the study has certain limitations. The analysis remains correlational, making it difficult to establish causal links between IRC and research performance. Future research could incorporate longitudinal studies and subject-specific analyses to refine these metrics further. Additionally, examining external factors such as funding, institutional size, and subject area specializations could improve our understanding of collaboration dynamics. Further, the work has demonstrated computation of the proposed indicators on data downloaded from Dimensions database, a major reason being the larger coverage of Dimensions database (Singh et al., 2021). However, these values for the institutions may vary if data from a different database is used. In this sense, the proposed indicators, like all the bibliometric indicators in existence, are also sensitive to the database used, and indicator quality will also be related to the quality of the database.

By introducing a structured framework to evaluate collaboration's impact, this work provides a valuable perspective on institutional research productivity. The boost formalism offers a scalable and robust model for assessing the effectiveness of international collaborations, guiding institutions and policymakers toward data-driven research strategies.

Declarations

Funding and/or Conflicts of interests/Competing interests: The authors declare that no funding was received for this work. Further, the manuscript complies with the ethical standards of the conference and there is no conflict of interest whatsoever.

References

Abramo, G., DÁngelo, A.C. & Murgia, G. (2017). The relationship among research productivity, research collaboration, and their determinants. *Journal of Informetrics*, 11(4), 1016-1030.

Abramo, G., DÁngelo, A.C. & Solazzi, M. (2011). The relationship between scientists' research performance and the degree of internationalization of their research. *Scientometrics*, 86(3), 629-643.

Adams, J. (2012). "Collaborations: The rise of research networks." *Nature*, 490(7420), 335–337. DOI: 10.1038/490335a.

Braun, T., Glänzel, W., & Schubert, A. (1991). The bibliometric assessment of UK scientific performance—some comments on Martin's "reply". *Scientometrics, 20*(2), 359-362.

Beaver, D. D. (2001). Reflections on scientific collaboration, (and its study): past, present, and future. *Scientometrics, 52*(3), 365-377. doi:10.1023/a:1014254214337

Birnholtz, J. P. (2007). When do researchers collaborate? Toward a model of collaboration propensity. *Journal of the American Society for Information Science and Technology*, 58(14), 2226-2239.doi:10.1002/asi.20684

Boekholt, P., Edler, J., Cunningham, P., & Flanagan, K. (2009). Drivers of international collaboration in research. *European Commission*. https://doi.org/10.2777/81914

Bozeman, B., & Boardman, C. (2014). *Research collaboration and team science: A state-of-the-art review and agenda* (Vol. 17). New York: Springer.

Chinchilla-Rodríguez, Z., Bu, Y., Robinson-García, N., & Sugimoto, C. R. (2021). An empirical review of the different variants of the probabilistic affinity index as applied to scientific collaboration. *Scientometrics*, *126*, 1775-1795.

Davidson Frame, J., & Carpenter, M. P. (1979). International research collaboration. *Social studies of science*, *9*(4), 481-497.

D'Ippolito, B., & Ruling, C. C. (2019). Research collaboration in Large Scale Research Infrastructures: Collaboration types and policy implications. *Research Policy*, 48(5), 1282-1296. doi:10.1016/j.respol.2019.01.011

Dua, J., Singh, V. K., & Lathabai, H. H. (2023). Measuring and characterizing international collaboration patterns in Indian scientific research. Scientometrics, 128(9), 5081-5116.

Frandsen, T. F., & Nicolaisen, J. (2010). What is in a name? Credit assignment practices in different disciplines. *Journal of Informetrics*, *4*(4), 608–617.

Fuchs, J. E., Sivertsen, G., & Rousseau, R. (2021). Measuring the relative intensity of collaboration within a network. *Scientometrics*, *126*(10), 8673-8682.

Gauffriau, M. (2017). A categorization of arguments for counting methods for publication and citation indicators. Journal of Informetrics, 11(3), 672-684.

Glänzel, W. (2001). National characteristics in international scientific co-authorship relations. *Scientometrics*, *51*(1), pp.69-115.

Glänzel, W., & Schubert, A. (2001). Double effort= double impact? A critical view at international co-authorship in chemistry. *Scientometrics*, *50*(2), 199-214.

Harsanyi, M. A. (1993). Multiple authors, multiple problems—bibliometrics and the study of scholarly collaboration: A literature review. *Library and Information Science Research*, 15, 325–354

Inglesi-Lotz, R., & Pouris, A. (2013). The influence of scientific research output of academics on economic growth in South Africa: an autoregressive distributed lag (ARDL) application. *Scientometrics*, *95*, 129-139.

Katz, J.S. & Martin, B.R. (1997). What is research collaboration? *Research policy*, *26*(1), pp.1-18.

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). "Team size matters: Collaboration and scientific impact since 1900." *Journal of the Association for Information Science and Technology*, 66(7), 1323–1332. DOI: 10.1002/asi.23266.

Leclerc, M., & Gagné, J. (1994). International scientific cooperation: The continentalization of science. *Scientometrics*, *31*(3), 261-292.

Lee, S., & Bozeman, B. (2005). The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science*, 35(5), 673-702.

Leydesdorff, L., & Wagner, C. S. (2008). International collaboration in science and the formation of a core group. *Journal of Informetrics*, *2*(4), 317–325.

Lindsey, D. (1980). Production and citation measures in the sociology of science: the problem of multiple authorship. Social Studies of Science, 10, 145–162

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology, & Human Values*, *17*(1), 101-126.

Luukkonen, T., Tijssen, R., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics*, *28*(1), 15-36.

Mattsson, P., Laget, P., Nilsson, A., & Sundberg, C. J. (2008). Intra-EU vs. extra-EU scientific co-publication patterns in EU. *Scientometrics*, *75*(3), 555-574.

Ntuli, H., Inglesi-Lotz, R., Chang, T. Y., & Pouris, A. (2015). Does research output cause economic growth or vice versa? Evidence from 34 OECD countries. Journal of the Association for Information Science and Technology, 66(8), 1709–1716

Ochiai, A. (1957). Zoogeographical studies on the solenoid fishes found in Japan and its neighboring Regions - II. *Bulletin of the Japanese Society of Scientific Fisheries, 22*(9), 526-530.

Okubo, Y., Miquel, J. F., Frigoletto, L., & Doré, J. C. (1992). Structure of international collaboration in science: Typology of countries through multivariate techniques using a link indicator. *Scientometrics, 25*, 321-351.

Persson, O., Glänzel, W. & Danell, R. Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics* 60, 421–432 (2004). https://doi.org/10.1023/B:SCIE.0000034384.35498.7d

Price, D. J. de Solla. (1969). Measuring the size of science. *Proceedings of Israel Academy of Sciences and Humanities*

Rousseau, R. (2021). Going back in time: Understanding patterns of international scientific collaboration. *Journal of Scientometric Research, 10*(1), 126-129.

Salton G., McGill M.J., *Introduction to Modern Information Retrieval*, McGraw-Hill, N.Y. 1983

Schubert, A., & Braun, T. (1990). International collaboration in the sciences 1981–1985. *Scientometrics, 19*(1-2), 3-10.

Schubert, A., & Glänzel, W. (2006). Cross-national preference in co-authorship, references and citations. *Scientometrics, 69*, 409-428.

Singh, V. K., Nandy, A., Singh, P., Karmakar, M., Singh, A., Lathabai, H. H., ... & Kanaujia, A. (2022). Indian Science Reports: a web-based scientometric portal for mapping Indian research competencies at overall and institutional levels. *Scientometrics*, 127(7), 4227-4236.

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. Scientometrics, 126, 5113-5142.

van Eck, N. J., & Waltman, L. (2009). How to normalize co-occurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology, 60*(8), 1635–1651.

Wagner, C. S., Brahmakulam, I., Jackson, B., Wong, A., & Yoda, T. (2001). Science and technology collaboration: Building capacity in developing countries. *Document Number MR-1357.0-WB, RAND Corporation, Santa Monica, CA.*

Wagner, C. S., & Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34(10), 1608–1618.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*(2), 365-391

Yamashita, Y., & Okubo, Y. (2006). Patterns of scientific collaboration between Japan and France: Inter-sectoral analysis using Probabilistic Partnership Index (PPI). *Scientometrics, 68*(2), 303-324.

Zhou, Q., & Leydesdorff, L. (2016). The normalization of occurrence and Co-occurrence matrices in bibliometrics using Cosine similarities and O chiai coefficients. *Journal of the Association for Information Science and Technology, 67*(11), 2805-2814.

Zitt, M., Bassecoulard, E., & Okubo, Y. (2000). Shadows of the past in international cooperation: Collaboration profiles of the top five producers of science. *Scientometrics, 47*(3), 627-657.

https://doi.org/10.51408/issi2025_006

Bridging Classification Systems: The Potentialities of Artificial Intelligence in Developing Concordance Tables for Science, Technology, and Policy

Guendalina Capece[1], Cinzia Daraio[2], Flavia Di Costa[3]

[1] *guendalina.capece@unimercatorum.it*
Department of Engineering and Sciences, Universitas Mercatorum, Rome (Italy)

[2] *daraio@diag.uniroma1.it*
DIAG Sapienza University of Rome, Via Ariosto, 25, 00185 Rome (Italy)

[3] *flavia.dicosta@unimercatorum.it*
Department of Humanities and Social Sciences, Universitas Mercatorum, Rome (Italy)

Abstract

This paper explores the challenges and methodologies involved in aligning diverse subject classification systems through the development of concordance tables. It investigates prior efforts, identifies successful implementations, and evaluates employed methods. Using a multi-method approach, the research combines a literature review with Artificial Intelligence (AI)-enhanced content analysis in Scopus to identify trends and gaps in existing studies. The findings highlight the potential of AI-driven methodologies to improve automation and reliability in creating concordance tables while identifying areas for future research. The study emphasizes the importance and the limits of using AI for integrating classification systems, supporting knowledge organization, and facilitating science and innovation policy decision-making.

Introduction and research questions

As the pursuit of interdisciplinary research frequently encounters diverse and complex systems of knowledge classification, the challenge of aligning these disparate systems becomes increasingly significant. This paper delves into the intricate task of harmonizing various subject classification frameworks by developing concordance tables. By examining prior efforts and successful implementations, while also evaluating the methods employed, this study offers a comprehensive review using a multi-method approach.

Patent data, rich in technological details, have been crucial in showcasing the technological composition of industries (Griliches, 1990). The classification systems used by patent authorities provide high-resolution and hierarchical structures, essential for systematically linking technologies and industries for research purposes (Lafond and Kim, 2019). Historically, patent data have measured technological changes within industries through citation-weighted patent counts. While most changes are incremental and hard to detect without considerable technological shocks, advancements in data collection, natural language processing, and network analysis have introduced new indices to capture gradual technological shifts within industries (Kelly et al., 2018).

Regarding the importance of subject classification systems in informetrics, the key topics include:

1) *Scientific, literature-based*: Systems like Scopus and Web of Science (WoS) fall into this category, often referred to as paper classification.
2) *Technical, patent-based*: The International Patent Classification (IPC) is an example here, based on prior art classification.
3) *Industry sector-based*: This category organizes subjects according to various industrial sectors.

Connections can be established between the different types of classification. For instance: citations in patents to scientific literature may create a link between a patent classification and paper classifications; industries as funders of research papers may establish a link between industrial sector and paper classifications. Industries as assignees of patents may create a link between industrial sector and patent classifications. This study focuses on subject classification systems and aims to systematically analyse which attempts have been made to develop concordance tables between different subject classifications; which concordance tables have actually been created; which methods were used to create these and how successful these methods were, in terms of the degree of validity of the proposed concordance.

Studies on the science-technology-industry interface are confronted with the need to create concordance tables between technology (patent) and industry subject classifications (Schmoch et al., 2003; Schmoch, 2008; Lybbert & Zolas, 2014; Dorner, & Harhoff, 2018; Neuhäusler, Frietsch & Kroll, 2019; Goldschlag, Lybbert, & Zolas, 2020).

Goldschlag, Lybbert & Zolas (2020) applied a probabilistic linkage methodology, pioneered by Lybbert and Zolas (2014), to create concordances between USPC and CPC technology codes and various industry and product classifications, including the International Standard Industrial Classification (ISIC), the North American Industrial Classification System (NAICS), the Standard International Trade Classification (SITC), and the Harmonized System (HS) product codes. Utilizing these concordances, the analysis examined how technology-industry relationships evolved over time by allowing the set of contributing patents to vary. Findings revealed that the link between technologies and industries showed remarkable persistence, with a recent increase in the rate of change after decades of decline. Additionally, the research provided suggestive evidence demonstrating the economic relevance of the measure of technological change. Changes in the industry-technology composition were correlated with shifts in occupational composition, aligning with existing literature on the labor market effects of new technologies.

Neuhäusler, Frietsch & Kroll (2019) enhanced the probabilistic concordance between industry sectors and technology fields, building on prior work (Neuhäusler et al., 2017) by reallocating patents to industry sectors and expanding the database. The analysis further extended to a concordance between scientific disciplines and technology fields. The paper provided valuable insights into the nexus between technological and scientific outputs and economic sectors, building on previous research by Frietsch et al. (2017) and Neuhäusler et al. (2017). This study employed probabilistic concordances at the micro level, linking patents to industry sectors and

publications to technology fields. This method aggregated patents and publications into matrices of patent shares per technology field and sector, and publication shares per discipline and technology field.

Subject classification systems are essential for organizing and accessing knowledge across scientific, technological, and economic domains. However, the coexistence of multiple classification systems often creates challenges in ensuring consistency and interoperability. This paper systematically examines the development and implementation of concordance tables designed to align these diverse systems. Specifically, we address the following research questions: (1) What attempts have been made to develop concordance tables between various subject classifications? (2) Which concordance tables have been successfully created? (3) What methods have been employed in their development? (4) How effective are these methods in terms of the validity, accuracy, and utility of the proposed concordance? (5) How artificial intelligence (AI) could help in developing, maintaining and updating concordance tables?

The paper is organized as follows. The next section outlines the history of concordance tables up to its latest development. The subsequent section describes how Scopus AI can be useful in completing a selected review. The following section presents the potential usefulness of AI in developing, maintaining and updating concordance tables and the last section concludes the paper by highlighting its policy relevance.

Methods

We employ a multi-method approach combining a comprehensive literature review with AI-enhanced content analysis in Scopus. Scopus AI is used to identify patterns, trends, and gaps in existing studies.

An analytical overview of the key characteristics highlights the sophisticated functionalities of the system as listed below:

1) Enhanced robustness through advanced content analysis: By employing Scopus AI, our multi-method approach becomes significantly more robust. The tool's sophisticated algorithms provide detailed insights by analyzing the vast volumes of documents available in the Scopus database. This analysis aids researchers in positioning their own work effectively within the current academic landscape.
2) Identification of patterns and trends: Scopus AI can recognize and highlight recurring patterns and emerging trends within academic publications, offering researchers a thorough understanding of the latest developments and gaps in their field. This facilitates the identification of research opportunities and helps in crafting more relevant and impactful studies.
3) Comprehensive literature review: By combining traditional literature review methods with AI-enhanced analysis, researchers can streamline their review process. Scopus AI quickly processes and categorizes data, ensuring a more exhaustive and precise literature review.

4) Data-driven insights: Scopus AI provides researchers with data-driven insights and analytics, helping them to make informed decisions about their research direction and focus areas. This data can be pivotal in identifying under-researched topics or confirming the significance of ongoing studies.
5) Efficiency and accuracy: The use of AI in content analysis significantly reduces the time and effort required to sift through massive amounts of literature while increasing accuracy. Researchers can rely on Scopus AI to update them with the most relevant and recent publications in their domain.

Results

State of the art on concordance tables

The requirements of a classification system, regardless of the specific application area (Fettke and Loos, 2003) are listed below:

- *Completeness*: the specific application domain should be completely covered by the classification scheme.
- *Precision*: a classification scheme must describe models at different levels of detail. The precision of a classification can be increased by defining new classes narrower.
- *Consistency*: the classification scheme must be free from contradictions.
- *Extensibility*: a classification system should be extensible so that they can be adopted in the future. It has extensibility if the new classification characteristics remain stable even after the addition or removal of some classes from the scheme.
- *User-friendliness*: classification scheme should be clearly understood.
- *Economic efficiency*: different type of costs for development and implementation of classification system.

The aim of classification is ordering entities into groups or classes based on their similarity: that means, from a statistical point of view, trying to minimize within-group variance and maximizing the between-group variance. Consequently, it seeks to realize groups that are as different (non-overlapping) possible with the maximum degree of similarity within each group. The basic rule of classification is set up classes that are both exhaustive and mutually exclusive. Typology is another term for a classification: it is multi-dimensional and conceptual. Taxonomy is a term similar to Typology, used as synonym, although it ought to be preferably used for classification of empirical entity (Bayley, 1994).
Listed in the

Table are pros and cons of classification schemes.

Table 1. Classification schemes - pros and cons.

PROS	*CONS*
It is a descriptive tool	Classification is descriptive, pre-explanatory or non-explanatory
Reduction of complexity that allows to synthesize a large amount of data in a smaller number of Types (taxa) significant	Static classification
Identification of the similarities and/or identification of differences in a complementary manner	Difficulty to choose the size and finding cases for classification
Submission of an exhaustive list of dimensions	The logic of the classes because typologies are criticized as dependent on the logic of classes rather than the use of continuous date as in the modern statistical techniques.
Comparison of Types	Although the types are often purely descriptive, however, they serve for the study of relationships and also for the specification of hypotheses concerning these relationships

1. Patent Classification

Internationally, the classification system is the International Patent Classification (IPC) which is updated periodically. The IPC was established in 1971 by the Strasbourg Agreement to provide and ensure a harmonized, hierarchical system for classifying the technology contained in patents and utility models. The current version of the IPC (2022) divides technology into eight sections (A-H) with approximately 75,000 subdivisions. According to the last version of IPC guide (2022):

"The Classification, being a means for obtaining an internationally uniform classification of patent documents has as its primary purpose the establishment of an effective search tool for the retrieval of patent documents by intellectual property offices and other users, in order to establish the novelty and evaluate the inventive step or non-obviousness (including the assessment of technical advance and useful results or utility) of technical disclosures in patent applications".

The Classification, furthermore, has the important purposes of serving as:

a) an instrument for the orderly arrangement of patent documents in order to facilitate access to the technological and legal information contained therein;

b) a basis for selective dissemination of information to all users of patent information;
c) a basis for investigating the state of the art in given fields of technology;
d) a basis for the preparation of industrial property statistics which in turn permit the assessment of technological development in various areas".

"The IPC is a hierarchical classification system. The contents of lower hierarchical levels are subdivisions of the contents of the higher hierarchical levels to which the lower levels are subordinated. The Classification separates the whole body of technical knowledge using the hierarchical levels, i.e., section, class, subclass, group and subgroup, in descending order of hierarchy".

The patenting system has a classification problem. The current classification system is based on technological and functional principles. The classification scheme is built from a technical point of view: an invention is normally classified according to its function or intrinsic nature.

According to IPC Guide (2022) page 22:

"As an application-oriented reference usually points from a function-oriented place to an application-oriented place, so an informative reference usually points from an application-oriented place to a function-oriented place".

"When it is unclear whether to classify a technical subject in a function-oriented place or in an application-oriented place, the following should be observed:

a) If a particular application is mentioned, but not specifically disclosed or fully identified, classification is made in the function-oriented place, if available. This is likely to be the case when several applications are broadly stated.
b) If the essential technical characteristics of the subject relate both to the intrinsic nature or function of a thing and to its particular use, or its special adaptation to or incorporation into a larger system, classification is made in both the function-oriented place and the application-oriented place, if available.
c) If guidance indicated in subparagraphs (a) and (b), above, cannot be used, classification is made in both the function-oriented place and the relevant application-oriented places".

2. Industry sectors Classification (IPC-industry concordances)

Industry classification or industry taxonomy organizes companies into industrial groupings based on similar production processes, similar products, or similar behavior in financial markets. A wide variety of taxonomies is in use, sponsored by different organizations and based on different criteria (Table 2).

Table 2. Industry classifications
Industry classification. Retrieved 08:29, January 21, 2025, from https://en.wikipedia.org/w/index.php?title=Industry_classification&oldid=1220947550.

ABBREVIATION	*FULL NAME*	*SPONSOR*	*CRITERION/ UNIT*	*NODE COUNT BY LEVEL*	*ISSUED*
ANZSIC	Australian and New Zealand Standard Industrial Classification	Governments of Australia and New Zealand			1993, 2006
BICS	Bloomberg Industry Classification Standard[2]	Bloomberg L.P.		10/.../2294	
GICS	Global Industry Classification Standard	Standard & Poor's, MSCI	market/ company	2-8 digits 11/24/69/158	1999–present (2018)
HSICS	Hang Seng Industry Classification System[3]	Hang Seng Indexes Company	Revenue source	11/31/89	
IBBICS	Industry Building Blocks[4]	Industry Building Blocks	Market line of business	19/130/550/3000/20200	2002
ICB	Industry Classification Benchmark	FTSE	market/ company	11/20/45/173	2005–present (2019)
ISIC	International Standard Industrial Classification of All Economic Activities	United Nations Statistics Division	production/ establishment	4 digits 21/88/238/419	1948–present (Rev. 4, 2008)
MGECS	Morningstar Global Equity Classificati	Morningstar, Inc.	Securities behavior	3/14/69/148	

ABBREVIATION	*FULL NAME*	*SPONSOR*	*CRITERION/ UNIT*	*NODE COUNT BY LEVEL*	*ISSUED*
	on System[5]				
NACE	Statistical Classification of Economic Activities in the European Community	European Union	production/ establishment	6 digits	1970, 1990, 2006, 2023
NAICS	North American Industry Classification System	Governments of the United States, Canada, and Mexico	production/ establishment	6 digits 17/99/313/724/1175 (/19745)[1]	1997, 2002, 2012, 2017, 2022
RBICS	FactSet Revere Business Industry Classification System	FactSet, acquired in 2013[6]	line of business	11000	
SIC	Standard Industrial Classification	Government of the United States	production/ establishment	4 digits 1004 categories	1937–1987 (superseded by NAICS, but still used in some applications)
SNI	Swedish Standard Industrial Classification	Government of Sweden			
TRBC	The Refinitiv Business Classification	Refinitiv	market/ company	10 digits 13/33/62/154/898[7]	2004, 2008, 2012, 2020[8]

ABBREVIATION	*FULL NAME*	*SPONSOR*	*CRITERION/ UNIT*	*NODE COUNT BY LEVEL*	*ISSUED*
UKSIC	United Kingdom Standard Industrial Classification of Economic Activities	Government of the United Kingdom			1948–present (2007)
UNSPSC	United Nations Standard Products and Services Code	United Nations	Product	8 digits (optional 9th) (four levels)	1998–present

The patent classification (IPC) and the industrial classification are not directly comparable. Three are the criteria for assigning an invention to an industry:

1) origin based: patents are assigned to the industrial sector of origin (the main economic sector of inventing / applicant company) (industry of origin);
2) producer based;
3) user based: patents are assigned to the sector where it is in use (the main industry to which belongs the product incorporating the invention) (industry of destination or industry of use).

There are three levels at which patents can be linked to economic activity:

1) macro-level (country): for study rate of innovation, country's innovative capacity effects of patent harmonization;
2) meso level (industry): for studying relationship between patenting and economic activity through time, space and technological classes;
3) micro-level (firm): patenting as part of firm-level strategies.

At meso level the link between patent and industry is based on concordance tables.

3. Concordances

Over the past decades, several notable concordance tables have been developed to map different classification systems (e.g., patent classifications to industry or product codes). Rather than describing each approach in detail within the text, we summarize the main characteristics of these concordances in Table 3. The table highlights the year, classification systems, methodology (e.g., probabilistic vs. direct mappings), and principal contributions of each notable concordance effort.
This consolidated view underscores the diverse methodological approaches—from manual mapping of IPC subclasses to industrial codes (Schmoch et al., 2003) to

algorithmic linkages using textual descriptions (Lybbert & Zolas, 2014)—and reveals how each method addresses particular research needs. It also illustrates how concordances have gradually become more probabilistic and data-driven, reflecting broader trends in AI and big data analytics.

By presenting these concordances side by side, we provide readers with a straightforward means to compare their strengths, limitations, and contexts of application (e.g., macro-level policy analysis vs. micro-level firm strategy). We refer to specific details of each study only when needed for interpreting our results, thus avoiding repetitive textual descriptions in the main body.

Table 3. Comparative Overview of Key Concordance Tables.

Concordance	Year	Classification Systems Mapped	Mapping Method	Key Contribution/Notes
Yale Technology Concordance (YTC) (Evenson et al., 1991)	1991	Patent (Canadian) → Industry (IOO & IUO)	Probabilistic/Manual	Early effort linking patents to industry of origin/use
DG Concordance (Schmoch et al., 2003)	2003	IPC → ISIC (44 sectors)	Manual mapping	Widely used in patent statistics; basis for Eurostat
ALP (Lybbert & Zolas, 2014)	2014	IPC → ISIC/SITC	Algorithmic/probabilistic	Introduces textual keywords & Bayesian weighting
Inventor-Establishment (Dorner & Harhoff) (2018)	2018	Patent (EPO) → Industry (NACE)	Micro-level matching	Leverages inventor-level data for higher precision
Others (OECD, MERIT, etc.)	Various	Patent → Industry/Product classifications	Mixed methods	Show incremental improvements & expansions

Notes:

- *IPC: International Patent Classification; ISIC: International Standard Industrial Classification; SITC: Standard International Trade Classification; NACE: Statistical Classification of Economic Activities in the European Community.*
- *IOO/IUO: Industry of Origin/Industry of Use.*
- *The approaches vary in granularity (e.g., macro-level vs. micro-level) and complexity (manual vs. algorithmic).*

The few studies whose goal was to design a concordance between industry sectors and technology classifications are listed in the following table (Table 4).

Table 4. Articles with concordance tables.

Concordance scheme	*Content*
OTAF Concordance (1974)	A computerized method has been developed by the OTAF at the U.S. Patent and Trademark Office to establish links based on a concordance of detailed patent classification codes and industry codes.
Evenson, R. E., Putnam, J. & Kortum, S. (1991)	Concordance table based on the industry classification made by the Canadian Intellectual Property Office that assigned both an industry of origin code (IOO) and an industry use code (IUO) to Canadian patents.
Kortum, S., & Putnam, J. (1997)	YCT developed within a probabilistic framework, linking potential industries. A statistical model predicts industry standard errors.
Verspagen, B., Morgastel, T. v., Slabbers, M. (1994)	MERIT Concordance matches IPC subclasses to 22 industrial classes based on a mix of 2- and 3-digit ISIC codes.
Johnson, D. K. (2002)	The OECD Technology Concordance (OTC), similar to the Yale Technology Concordance, serves as an instrument for converting IPC-based patent data into patent counts categorized by economic sector.
Schmoch, U., Laville, F., Patel, P. & Frietsch, R. (2003)	The "DG Concordance" aims to align IPC subclasses with ISIC industry classifications, assigning 625 IPC subclasses to 44 manufacturing sectors, each associated with one or more ISIC codes.
Lybbert, T. J., & Zolas, N. J. (2014)	The "Algorithmic Links with Probabilities" (ALP) approach constructs concordances between the IPC system and industry classification systems like SITC and ISIC. It uses keywords from industry descriptions and a probabilistic framework to match data, providing meso-level mappings that complement macro- and firm-level mappings.

Concordance scheme	*Content*
van Looy, B., Vereyen, C., & Schmoch, U. (2014)	The concordance update, addressing the limitations of the 2003 Schmoch et al. version, widely utilized by Eurostat for patent statistics, reviewed 44 technology definitions, assigned IPC 4-digit codes, and incorporated the NACE 2 classification.
Dorner, M., & Harhoff, D. (2018)	Concordance tables, address the preference for linking industries to their knowledge and technological opportunities ("industry of origin") using linked inventor-establishment data for Germany to generate accurate industry of origin information for patents.
Neuhäusler, P., Frietsch, R., & Kroll, H. (2019)	Concordance tables integrate micro-level data from patent applicants and authors, aggregated at sector and technology field levels, utilizing sources like NACE Rev. 2 and the 35 WIPO fields, and applying probabilistic methods to generate comprehensive concordances.
Goldschlag, N., Lybbert, T. J., & Zolas, N. J. (2019)	Concordance tables, using a probabilistic linkage methodology, were created to map USPC and CPC technology codes to various industry and product classifications, including ISIC, NAICS, SITC, and HS.

An application of Scopus AI

The following steps outline the procedure for effectively utilizing Scopus AI in research, ensuring a comprehensive and data-driven understanding of the field (Table 5):

Table 5. Scopus AI procedure.

Step	*Description*
1. Accessing Scopus AI	Start by logging into the Scopus account, ensuring that the institution has access to the Scopus AI features.
2. Formulating Queries	Natural language is used to type questions or statements into the Scopus AI search box. No need for complex search strings.
3. AI-Enhanced Search	Sophisticated algorithms analyze the vast volume of documents, identifying patterns, trends, and knowledge gaps.
4. Reviewing Topic Summary	Scopus AI synthesizes abstracts from relevant documents to generate a Topic Summary, offering an overarching view of the subject.
5. Exploring Expanded Summary	Detailed information is provided through the Expanded Summary, offering comprehensive insights and references to supporting documents.
6. Utilizing Concept Maps	Keywords from research abstracts are used to generate an interactive Concept Map, illustrating the relationship with various subtopics.
7. Identifying Key Papers and Researchers	Scopus AI highlights influential papers and top researchers linked to the query, helping identify critical publications and leading experts.
8. Data-Driven Insights	Insights and analytics provided by Scopus AI are utilized to make informed decisions about research direction, identifying under-researched areas.
9. Saving and Exporting Results	After reviewing, results can be saved and exported for further analysis and reference. Summaries and insights can be easily shared with the research team.

To test the potentialities of the Scopus-AI tool, bibliographical research was conducted as described below.

A sentence containing the keywords (concordance table, subject classification, science classification, patent classification, industry classification) from Daraio, Di Costa & Moed (2014) was entered into the search box.

Input in search box

"How can concordance tables facilitate the alignment of different subject classification, science classification, patent classification, and industry classification systems"

The Scopus -AI hidden procedure that generates the results is the following:

1) **Creating a plan to answer your query**
2) **Performing natural language search:**
   How do concordance tables facilitate the alignment of different classification systems in subjects, sciences, patents, and industries?

3) **Performing keyword search:**
("concordance" OR "alignment" OR "correspondence" OR "mapping") AND ("tables" OR "charts" OR "lists" OR "diagrams") AND ("classification" OR "categorization" OR "taxonomy" OR "system") AND ("systems" OR "frameworks" OR "models" OR "structures")
4) **Generating summary**

Scopus-AI creates a structured summary in the following main sections, accompanied by the respective references:

1) How Concordance Tables Facilitate Alignment of Different Classification Systems
2) Linking Different Classifications
3) Policy Planning and Technology Targeting
4) Empirical Innovation Studies
5) Advantages and Applications
6) Challenges and Considerations
7) Conclusion

Scopus AI utilizes keywords extracted from research abstracts to generate interactive Concept Maps. This tool offers a comprehensive overview of the topic landscape, illustrating connections with various research areas, including those that may extend beyond the user's traditional scope of inquiry. Figure 1 displays the results of the example under consideration.

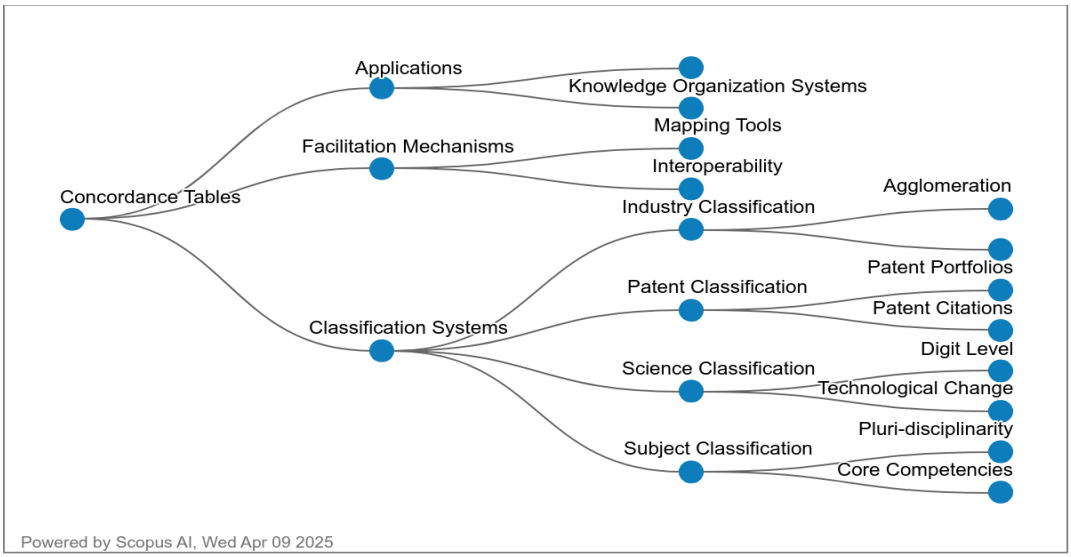


Figure 1. Concept Map generated by Scopus AI on Apr 09 2025.

Benefits and challenges

Concordance tables offer numerous benefits, guiding policymakers in understanding the gaps between science, technology, and industry, aiding in targeted policy planning and technology development (Wong, & Fung, 2017). They facilitate the utilization of often-underutilized patent documents and technical information, enabling the visualization and analysis of relationships among technologies, which supports more informed decision-making (Pasek, 2021; Leydesdorff, 2008). By highlighting connections between different fields, concordance tables foster innovation by identifying opportunities for cross-disciplinary research and development, helping track the evolution of technologies and industries, and providing a roadmap for future innovation (Lee, 2018; Wong, & Fung, 2017).

However, several challenges accompany concordance tables. The alignment of different classification systems like IPC and CPC, each with distinct logic and granularity, is complex and often leads to misalignment and oversight of emerging technological trends (Lobo, & Strumsky, 2019; Alisova, 2013). Aggregating bibliographic data from diverse sources poses technical difficulties, requiring high-quality mappings to resolve defects such as missing or incorrect relations (Pfeffer, 2016; Ivanova & Lambrix, 2013). Automated methods, while less resource-intensive, may not achieve the same level of accuracy (Pfeffer, 2016). Despite these challenges, concordance tables remain a valuable tool at the intersection of technology and industry. See Table 6 for a summary of the main challenges and benefits of concordance tables.

Table 6. Challenges and benefits of concordance tables to align classification systems.

Challenges	*Benefits*
Complexity and diversity of systems (Lobo, & Strumsky, 2019; Alisova, 2013; Lee, 2018)	Policy and planning support (Wong, & Fung, 2017)
Data integration and quality issues (Pfeffer, 2016; Ivanova & Lambrix, 2013)	Enhanced data utilization (Pasek, 2021; Leydesdorff, 2008)
Resource intensity (Pfeffer, 2016)	Support for innovation (Lee, 2018; Wong, & Fung, 2017)

Potentialities and limits of AI usage for concordance tables

AI-driven techniques provide solutions to several key challenges in mapping classification systems, including scalability, semantic ambiguity, and the need for dynamic updates. By combining deep learning for semantic understanding, clustering for pattern detection, and predictive modeling for adaptability, AI introduces a powerful set of tools to automate, refine, and expedite the concordance process. Moreover, the hybrid integration of AI with human expertise ensures that the benefits of automation are paired with contextual precision, resulting in robust and accurate mapping frameworks.

Artificial intelligence—specifically, natural language processing (NLP) and machine learning (ML)—be utilized to *automate, refine, accelerate* and *validate* the mapping process.

Natural language processing (NLP) can analyze and interpret textual descriptions, category names, and associated metadata in classification systems. Specific applications include i) *Semantic Analysis* in which NLP algorithms extract the meaning of terms and phrases from classification systems, identifying synonyms, hierarchical relationships, and contextual overlaps; ii) *Entity Recognition* in which NLP can identify and tag key concepts, entities, and terms from textual data, enabling precise alignment between systems; iii) *Text Clustering*: based on NLP-powered clustering that groups similar terms across classifications, revealing patterns of equivalence or correspondence.

Machine learning algorithms can improve the precision of concordance mapping through i) *Supervised Learning* through which ML models trained on labeled datasets can learn to map terms from one classification system to another, generalizing their knowledge to new, unseen classifications; ii) *Unsupervised Learning* based on techniques like clustering or topic modeling can identify hidden relationships in datasets without requiring pre-labeled data, making them ideal for exploratory concordance creation; iii) *Contextual Embeddings* based on advanced ML methods like transformer-based models that can embed terms and categories in high-dimensional spaces, enabling similarity detection based on context.

AI technologies significantly reduce the time and effort required for mapping by automating repetitive and computationally intensive tasks. Large datasets spanning multiple classification systems can be processed simultaneously, scaling concordance efforts beyond manual capabilities. AI systems can automatically update mappings as classification systems evolve or new data becomes available.

Finally, AI can enhance the reliability and validity of the mappings by error detection, and identifying inconsistencies or ambiguities in the concordance through anomaly detection algorithms.

AI-driven techniques—such as deep learning for semantic analysis, clustering for unsupervised categorization, and predictive modeling for trend analysis— could be incorporated to overcome the limitations of traditional tools with *hybrid methodologies*, based on the combination of manual expertise with automated AI processing. AI systems are powerful but not infallible. They are particularly proficient at handling large-scale, repetitive tasks, while human experts excel at nuanced judgment and contextual understanding. By combining AI-driven techniques with manual expertise, the limitations of both approaches can be mitigated.

Additionally, AI enables dynamic and adaptive concordance tables that evolve with new data inputs, reflecting real-time changes in classifications.

AI tools offer transformative opportunities for policymakers to make better use of concordance tables by enhancing their accessibility, adaptability, and utility in decision-making processes. By incorporating AI-driven techniques, concordance tables can be transformed from static tools into dynamic, interactive systems that provide real-time updates, visual analytics, and predictive insights.

AI can monitor data continuously updated, such as scientific publications and patents, to ensure concordance tables remain current and reflect the latest developments. This allows policymakers to make decisions based on the most up-to-date information, particularly in rapidly evolving fields. Furthermore, AI-powered visual analytics tools, such as dashboards and network graphs, can present complex concordance relationships in an intuitive and actionable format. For example, policymakers could use these tools to identify overlaps or gaps in innovation funding across sectors or to explore regional trends in research output.

Another critical capability of AI is its ability to provide predictive insights and scenario modeling. By simulating the potential outcomes of classification alignments, AI tools can help policymakers anticipate the effects of their decisions on various sectors, such as predicting the impact of aligning academic and industry classifications on workforce development or innovation growth. Moreover, these tools can be tailored to provide policy-specific recommendations, allowing policymakers to explore how concordance relationships affect their goals and constraints.

Despite these advantages, the use of AI in this domain is not without *significant limitations*. A major challenge lies in the dependence on the *quality* and *representativeness* of the training datasets. *Incomplete* or *biased* data can lead to inaccuracies in concordance mappings, perpetuating existing discrepancies rather than resolving them. Additionally, many AI models operate as opaque *black boxes*, where their processes and outputs are not easily interpretable. This *lack of transparency* can undermine trust among stakeholders and impede the reproducibility of results, a critical aspect of scientific rigor.

AI tools also face difficulties in capturing domain-specific distinctions, particularly in highly specialized or interdisciplinary fields. While AI excels at automating repetitive tasks, its ability to make contextually informed decisions is limited, *necessitating continued human supervision*. Furthermore, the dynamic nature of classification systems, while well-suited to AI's adaptive capabilities, introduces challenges in maintaining long-term consistency. Frequent updates to concordance tables can lead to *fragmentation* or misalignment of historical data.

The *ethical implications* of AI use are another pressing concern. *Bias* in AI models, if unchecked, can exacerbate existing misuses, and the use of proprietary or sensitive data raises questions about *privacy* and *intellectual property*. Infrastructure and expertise requirements present additional barriers, as deploying and maintaining sophisticated AI systems often demands significant computational resources and technical skills. These constraints can limit accessibility for smaller organizations or underfunded research initiatives, creating disparities in who can leverage these tools effectively. Moreover, overreliance on AI risks neglecting the critical evaluative role that human judgment plays in ensuring accuracy and relevance.

To address these challenges, the research community and policymakers must prioritize efforts to mitigate biases in training datasets and promote the development of transparent, interpretable AI models. Collaborative frameworks that bring together AI developers, domain experts, and decision-makers are essential to ensure that AI-driven concordance tools produce balanced and meaningful outcomes. The

creation of resource-efficient and cost-effective solutions is equally important to expand accessibility across diverse institutions. Ethical oversight and accountability mechanisms should be established to monitor AI usage, safeguard data privacy, and foster trust.

While the limitations of AI tools are significant, they do not compensate the transformative potential these technologies hold for enhancing the efficiency, precision, and adaptability of concordance tables. By addressing these issues thoughtfully, AI can serve as a powerful catalyst for aligning classification systems and advancing innovation policies in an increasingly interconnected world.

AI-driven techniques have greatly enhanced the creation and maintenance of concordance tables by automating key tasks such as large-scale text analysis, semantic clustering, and predictive modelling. Through natural language processing (NLP) and machine learning (ML) methods—including deep learning and transformer-based embeddings—AI can reduce manual effort, scale mapping efforts across large, diverse datasets, and dynamically update concordances in response to newly available information. In doing so, policymakers and researchers gain real-time insights, enabling more informed decisions about resource allocation, innovation funding, and strategic planning.

Despite its advantages, the effectiveness of AI depends heavily on high-quality, representative training data and transparent, interpretable models. Biased or incomplete datasets can reinforce existing discrepancies, while black-box approaches make it difficult to validate or reproduce results. AI also struggles with domain-specific nuances, requiring ongoing human supervision and expert input. Moreover, adopting sophisticated AI systems often demands significant computational resources, specialized expertise, and robust data governance—factors that may restrict access for smaller organizations. Ethical and legal considerations, such as bias, data privacy, and intellectual property, further complicate large-scale adoption.

Addressing these challenges calls for collaborative frameworks among AI developers, domain experts, and policymakers, alongside efforts to develop resource-efficient, explainable AI solutions. With proper oversight and strategies to mitigate biases, AI tools can serve as powerful catalysts for enhancing the efficiency, precision, and adaptability of concordance tables, promoting more effective alignment of classification systems in an increasingly interconnected world.

AI Techniques for Developing and Updating Concordance Tables

Recent advancements in Artificial Intelligence (AI)—particularly in Natural Language Processing (NLP) and Machine Learning (ML)—offer powerful tools to address the complexities of aligning diverse classification systems (e.g., patent, industry, and scientific taxonomies). In this context, AI can**:**

1. *Automate Large-Scale Text Analysis*. NLP methods such as named-entity recognition and text clustering enable the systematic extraction and grouping of relevant terms or codes from extensive document corpora (patents, scientific articles, etc.). These methods can detect semantic overlaps, synonyms, or

hierarchical relationships that inform how different classification systems interrelate.

2. *Improve Accuracy and Reduce Redundancies***.** By applying supervised learning (e.g., Random Forest, SVM, or neural networks) to labeled training sets, AI algorithms learn to associate the descriptive content of documents with specific industry or patent classes. This reduces time-consuming manual mapping and can facilitate ongoing updates as new data emerge.
3. *Identify Ambiguities and Cross-Disciplinary Links***.** NLP-driven topic modeling and clustering (e.g., LDA, DBSCAN) can discover hidden patterns and overlapping categories. This is particularly useful when dealing with interdisciplinary fields, where traditional classification schemes may lack clarity or granularity.
4. *Enable Dynamic, Scalable Concordance Tables***.** Machine learning approaches can update mappings in near real-time, reflecting evolving research frontiers or emerging technologies. Hybrid "human-in-the-loop" workflows, in which experts validate uncertain assignments, further enhance reliability and transparency.

Practical Examples

Works such as Lybbert and Zolas (2014) employ algorithmic text-matching to link International Patent Classification (IPC) codes with economic and industry classifications (e.g., ISIC, SITC). Similarly, Dorner and Harhoff (2018) leverage inventor-establishment data to refine the accuracy of patent-to-industry correspondences. Although in the previously cited approaches there are not AI techniques applied, these methods illustrate how AI-driven techniques can increase both the speed and precision of concordance-building efforts, helping policymakers and scholars navigate constantly evolving classification systems.

Addressing Ethical and Infrastructural Challenges

Despite these advantages, AI-based approaches also raise important ethical and infrastructural considerations (Table 7).

Table 7. AI-based approaches ethical and infrastructural considerations.

Challenges	*Description*
Bias in Training Data	Algorithmic decisions can inadvertently reflect biases in the underlying datasets, especially if certain industries, countries, or languages are underrepresented. Periodic audits and balanced data sampling can help reduce these distortions
Data Privacy and Confidentiality	Large-scale text analysis often involves sensitive corporate data, personal details (e.g., inventor information), or confidential product descriptions. Employing robust data governance strategies—such as anonymization protocols and secure

	storage—ensures compliance with legal and ethical standards
Interpretability and Accountability	Many advanced AI models (e.g., deep neural networks) operate as "black boxes," complicating the explanation of how specific concordances are generated. Solutions include explainable AI frameworks and transparent reporting of model decisions.
Infrastructure and Accessibility	Training and deploying AI models can require significant computational resources and specialized expertise. Smaller research groups or institutions may lack the necessary hardware, software, or funding to implement advanced methods, potentially widening the gap in data capabilities across organizations
Dynamic Maintenance Over Time	Because classification systems evolve, concordance tables must be continuously updated. AI can facilitate automated or semi-automated revision, but this introduces ongoing costs in software maintenance, model retraining, and data curation

By actively managing these *ethical and infrastructural challenges*, researchers and policymakers can maximize the benefits of AI-driven concordance mapping—greater speed, scalability, and accuracy—while ensuring fair, secure, and transparent processes.

Conclusions and further development

This study emphasizes the central role of concordance tables in harmonizing diverse classification systems across scientific, technological, and industrial domains. Through a detailed exploration of historical developments, methodological advancements, and the integration of Artificial Intelligence (AI), our research sheds light on the opportunities and challenges inherent in aligning these systems. Concordance tables are indispensable tools for fostering interoperability, facilitating knowledge organization, and supporting evidence-based decision-making in science and innovation policy.

The integration of AI into the creation and maintenance of concordance tables marks a significant step forward. AI-driven tools such as Scopus AI demonstrate transformative potential in this context, enhancing automation, precision, and scalability. For instance, Scopus AI's ability to analyze vast datasets, identify patterns, and generate concept maps provides researchers with a comprehensive understanding of classification relationships. By summarizing information from diverse sources, the tool reveals knowledge gaps and highlights emerging trends, enabling the development of concordance tables that remain relevant in an ever-evolving landscape. These capabilities were evident in the AI-generated concept

maps and summaries, which revealed the intricate connections between subject classifications, patent classifications, and industry frameworks.

Despite these advances, the application of AI in this domain faces notable limitations. AI tools depend heavily on the quality, diversity, and neutrality of training datasets. Inadequate or biased data can lead to inaccuracies in mappings, undermining the reliability of concordance tables. Additionally, the *black-box* nature of many AI models poses challenges for interpretability and transparency, complicating efforts to validate and trust their outputs. Domain-specific distinctions and the dynamic evolution of classification systems further complicate the process, as these require a combination of automated processing and expert judgment to address.

The ethical and infrastructural considerations associated with AI tools also warrant attention. Biases in AI models, if unchecked, can exacerbate systemic biases, while issues related to data privacy and intellectual property remain pressing concerns. The computational resources and expertise required to implement sophisticated AI systems often limit their accessibility to well-funded organizations, creating disparities across the research landscape.

To harness the full potential of AI while addressing its limitations, future research must focus on several key areas. First, ensuring transparency and fairness in AI methodologies is essential. This involves developing explainable AI models and employing diverse training datasets to mitigate biases. Second, collaborative efforts between AI developers, domain experts, and policymakers are necessary to balance the computational power of AI with the contextual precision of human oversight. Third, the creation of resource-efficient AI tools can enhance accessibility, enabling broader participation in the development of concordance tables.

The Scopus AI tool offers an indication into the future of AI-powered research, demonstrating how interactive visualizations and real-time data analysis can support decision-making. By aligning patent classifications with academic research and industry frameworks, these tools provide actionable insights into the innovation ecosystem. For example, identifying gaps between research activity and patent filings could inform targeted funding strategies or reveal emerging technologies requiring early support.

Building on these insights, we propose *five* recommendations to guide both future research and policy initiatives:

i) *Promote Open, Interoperable Datasets by* establishing standardized metadata protocols so that patent, scientific, and industry data can be more easily integrated and by encouraging data sharing across institutions and countries through open-access repositories, enabling the creation of more accurate and universally applicable concordance tables.

ii) Develop Transparent and Fair AI Tools by prioritizing explainable AI approaches that allow for auditing and improving algorithmic decisions and by adopting bias-mitigation strategies, including balanced sampling and periodic model audits, to ensure underrepresented fields or regions are adequately reflected.

iii) *Enhance Collaborative Frameworks* by fostering partnerships between domain experts, AI developers, and policymakers to combine technical expertise with contextual knowledge and by encouraging cross-sectoral working groups to refine AI methodologies and evaluate their impact on policy decisions.

iv) *Create Policy Incentives for Dynamic Concordances* by integrating human-in-the-loop governance in official guidelines, ensuring experts validate AI outputs for sensitive sectors and by supporting sustainable funding models to maintain and update concordance tables, reflecting changes in classification systems and emerging technologies.

v) *Strengthen Ethical and Legal Frameworks* by implementing data privacy regulations that protect sensitive information while allowing large-scale text analysis and by enforcing accountability mechanisms (e.g., impact assessments, review boards) for teams employing AI in classifying or mapping potentially sensitive data.

By adopting these recommendations, researchers and policymakers can collaboratively move toward more effective, equitable, and innovative concordance-building efforts. In conclusion, while AI introduces significant advancements in the development and maintenance of concordance tables, its successful implementation requires a careful balance between automation and human expertise. Addressing the ethical, technical, and infrastructural challenges associated with AI is crucial for realizing its full potential. By adopting hybrid approaches and fostering collaborative frameworks, concordance tables can evolve into dynamic tools that not only align classification systems but also drive innovation and policy development in a rapidly changing knowledge-based world.

Acknowledgments

This work builds upon and expands the research initiated in Daraio, Di Costa, and Moed (2014) and is dedicated to the memory of Henk Moed.

References

Alisova, N. V. (2013). Biomedical engineering in international patent classification. Biomedical Engineering, 47(3), 164-168.

Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques*. Sage Publications, Inc.

Daraio, C., Di Costa, F., & Moed, H. F. (2014). Towards Concordance Tables of Different Subject Classification Systems. A literature review with policy implications. STI 2014 Leiden, 132-135.

Dorner, M., & Harhoff, D. (2018). A novel technology-industry concordance table based on linked inventor-establishment data. *Research Policy*, 47(4), 768-781.

Evenson, R. E., Putnam, J. & Kortum, S. (1991). Estimating patent counts by industry using the Yale-Canada concordance. Final Report to the National Science Foundation.

Fettke, P., & Loos, P. (2003). Classification of reference models: a methodology and its application. *Information systems and e-business management*, *1*, 35-53.

Frietsch, R., Kladroba, A., Markianidou, P., Neuhäusler, P., Peter, V., Ravet, J., ... & Schneider, J. (2017). Final Report on the Collection of Patents and Business Indicators by Economic Sector: Societal Grand Challenges and Key Enabling Technologies Collection and Analysis of Private R&D Investment and Patent Data in Different Sectors, Thematic Areas and Societal Challenges. Luxembourg: Publications Office of the European Union, Luxembourg, Accessed 27 January 2025.

Goldschlag, N., Lybbert, T. J., & Zolas, N. J. (2020). Tracking the technological composition of industries with algorithmic patent concordances. *Economics of Innovation and New Technology*, *29*(6), 582-602.

Griliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey. *Journal of Economic Literature* 28 (4): 1661–1707.

World Intellectual Property Organization. (2023). Guide to the International Patent Classification (IPC 2022 edition). World Intellectual Property Organization. Available at: https://www.wipo.int/edocs/pubdocs/en/wipo-guide-ipc-2022-en-guide-to-the-international-patent-classification-2022.pdf

Ivanova, V., & Lambrix, P. (2013). A unified approach for aligning taxonomies and debugging taxonomies and their alignments. In The Semantic Web: Semantics and Big Data: 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings 10 (pp. 1-15). Springer Berlin Heidelberg.

Johnson, D. K. (2002). The OECD Technology Concordance (OTC): Patents by industry of manufacture and sector of use.

Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring technological innovation over the long run. *American Economic Review: Insights*, *3*(3), 303-320.

Kortum, S., & Putnam, J. (1997). Assigning patents to industries: tests of the Yale Technology Concordance. Economic Systems Research, 9(2), 161-176.

Lafond, F., & Kim, D. (2019). Long-run dynamics of the US patent classification system. *Journal of Evolutionary Economics*, *29*(2), 631-664.

Lee, H. (2018, October). Research on the impact of technology taxonomy for the tracking of technology convergence. In *2018 International Conference on Information and Communication Technology Convergence (ICTC)* (pp. 1452-1456). IEEE.

Leydesdorff, L. (2008). Patent classifications as indicators of intellectual organization. *Journal of the American Society for Information Science and Technology*, *59*(10), 1582-1597.

Lobo, J., & Strumsky, D. (2019). Sources of inventive novelty: two patent classification schemas, same story. *Scientometrics*, *120*(1), 19-37.

Lybbert, T. J., & Zolas, N. J. (2014). Getting patents and economic data to speak to each other: An 'algorithmic links with probabilities' approach for joint analyses of patenting and economic activity. *Research Policy*, 43(3), 530-542.

Neuhäusler, P., Frietsch, R., & Kroll, H. (2019). Probabilistic concordance schemes for the re-assignment of patents to economic sectors and scientific publications to technology fields (No. 60). Fraunhofer ISI Discussion Papers-Innovation Systems and Policy Analysis.

Pasek, J. E. (2021). Trends in bioengineering patents granted. *Biomedical Sciences Instrumentation*, 57(2), 61-73.

Neuhäusler, P., Frietsch, R., Mund, C., & Eckl, V. (2017). Identifying the technology profiles of R&D performing firms—a matching of R&D and patent data. *International Journal of Innovation and Technology Management*, *14*(01), 1740003. Trends in bioengineering patents granted. Biomedical Sciences Instrumentation, 57(2), 61-73

Office of Technology Assessment and Forecast. (1974). OTAF Concordance between U.S. Patent Classification and Standard Industrial Classification Systems.

Pfeffer M. (2016). Automatic creation of mappings between classification systems for bibliographic data. In Proceedings of the 2016 International Conference on Dublin Core and Metadata Applications (DCMI'16). Dublin Core Metadata Initiative, 75–84.

Putnam, J., & Evenson, R. E. (1994). Inter-sectoral technology flows: Estimates from a patent concordance with an application to Italy. Mimeograph, Yale University, New Haven, CT.

Schmoch, U. (2008). Concept of a Technology Classification for Country Comparisons. Final Report to the World Intellectual Property Office (WIPO), Karlsruhe: Fraunhofer ISI.

Schmoch, U., Laville, F., Patel, P., & Frietsch, R. (2003). Linking technology areas to industrial sectors. Final Report to the European Commission, DG Research.

van Looy, B., Vereyen, C., & Schmoch, U. (2014). Patent statistics: Concordance IPC V8–NACE rev. 2. Eurostat, Euopean Commission.

Verspagen, B., Moergastel, T. V., & Slabbers, M. (1994). MERIT concordance table: IPC-ISIC (rev. 2).

Verspagen, B. (1997). Measuring intersectoral technology spillovers: estimates from the European and US patent office databases. *Economic Systems Research*, 9(1), 47-65.

Wikipedia contributors. (2024, April 26). Industry classification. In Wikipedia, The Free Encyclopedia. Retrieved 08:29, January 21, 2025, from https://en.wikipedia.org/w/index.php?title=Industry_classification&oldid=1220947550

Wong, C. Y., & Fung, H. N. (2017). Science-technology-industry correlative indicators for policy targeting on emerging technologies: exploring the core competencies and promising industries of aspirant economies. *Scientometrics*, 111, 841-867.

https://doi.org/10.51408/issi2025_075

Characterizing Global Gender Gaps in STEM Using Facebook Data

Carolina Coimbra Vieira[1], Marisa Vasconcelos[2]

[1]*coimbravieira@demogr.mpg.de*
Max Planck Institute for Demographic Research (Germany)
Max Planck Institute for Software Systems (Germany)
Saarland University (Germany)

[2]*marisa.vasconcelos@gmail.com*
Universidade Federal de Minas Gerais (Brazil)

Abstract

Despite progress in addressing gender inequality in education and labor market, fields such as Science, Technology, Engineering, and Mathematics (STEM) remain far from achieving gender parity. The COVID-19 pandemic accelerated digital transformation and increased the demand for STEM professionals, yet gender disparities in the workforce persist. Monitoring these gaps globally is essential for understanding emerging trends and informing policy. However, traditional data sources are often limited by cost, scope, and availability. In this study, we explore Facebook Ads data as a scalable and timely alternative for assessing gender disparities in interest in STEM-related field. We analyze user data from 198 countries across 142 Facebook interests linked to both STEM and non-STEM college majors. Our findings reveal that, in most countries, more Facebook users self-report as female than male. Nevertheless, male users express greater interest in STEM majors — especially Engineering and Technology—while female users tend to show higher interest in Life Sciences and Mathematics. Furthermore, we observe that countries with a lower proportion of male users interested in college majors tend to perform better on official gender gap indicators based on survey data. These findings highlight the potential of social media data as a complementary resource for monitoring global disparities in education and career interests.

Introduction

Rapid technological advances have disrupted industries and the job market, amplifying the global demand for STEM – Science, Technology, Engineering, and Mathematics – professionals (UNESCO, 2021). The COVID-19 pandemic has accelerated digital transformation; however, the observed impact varies across cultures, ethnic groups, and genders. Despite growing initiatives that aim to promote inclusion, women remain underrepresented in many STEM-related occupations, especially in high-tech sectors (WEF, 2016). Diverse representation in STEM is essential not only for equity but also for fostering innovation and economic resilience. Prior research highlights that teams with greater gender diversity tend to perform better and drive stronger business outcomes (Forbes, 2018; Gompers & Kovvali, 2018).

Despite ongoing efforts to narrow the gender gap in STEM, some initiatives seem ineffective, and progress toward parity is slow. For greater impact, strategies must consider the cultural specificities of each country or region (WEF, 2021) as well as disciplinary differences. Gender gap statistics often mask disparities within STEM fields. In the U.S., for example, women earn over half of undergraduate degrees in

Biology, Chemistry, and Mathematics, but only 20% graduate from fields such as Computer Science, Engineering, and Physics (Cheryan et al., 2016; Munoz-Boudet & Revenga, 2017).

Most studies on gender gaps rely on surveys (Garcia-Holgado & Garcia-Penalvo, 2022; Tandrayen-Ragoobur & Gokulsing, 2021), which require significant time and financial resources. Also, these studies tend to focus on gender gaps in education and labor markets, with less emphasis on preferences. On a global scale, consistent data collection is difficult, and statistics for many countries, especially in the Global South, remain scarce. To address these limitations, we explore the use of social media data to assess gender balance across users' interests in STEM and non-STEM majors. In this paper, we present a large-scale analysis of the global STEM gender gap using data from Facebook Advertising Platform (Facebook Ads), where gender is self-reported by users in their profiles.

Throughout this paper, we use the terms *female/male users* to refer to individuals on Facebook who self-report their gender as female or male. While these labels reflect the binary options provided by the platform, they do not necessarily correspond to gender identity. We use *women/men* when referring to offline data sources or broader gender-related discussions to maintain consistency with those sources.

Facebook Ads data is widely used in various contexts, including assessing population health (Araujo et al., 2017), inferring political views (Guimarães et al., 2021), measuring cultural similarities (Vieira et al., 2022), predicting migration patterns (Alexander et al., 2019; Palotti et al., 2020), and conducting gender gap assessments (Garcia et al., 2018; Mejova et al., 2018; Vieira & Vasconcelos, 2021). Our study provides a global analysis of gender gaps in interest across a broad set of STEM and non-STEM interests associated with college majors.

As part of our methodology, we curated a list of Facebook interests associated with college majors in both STEM and non-STEM fields. Then, we collected the estimated number of Facebook users expressing interest in each major across multiple countries. To evaluate gender gaps, we derived two measures: the Overall Gender Balance (OGB) and the Gender Balance (GB). OGB represents the proportion of male Facebook users in a given country, while GB quantifies the proportion of male users in that country interested in a specific major. We applied these metrics to analyze variations in gender gaps across 142 college majors in 198 countries.

Our findings reveal a contrast between interests related to STEM and non-STEM majors. While Facebook users interested in non-STEM majors are predominately female, users interested in STEM majors are predominately male. However, within the STEM, differences emerge; for instance, Life Sciences and Math attract relatively more female users, whereas Engineering and Technology are more popular among male users. Non-STEM majors, such as Economics, Business, History, Government, and Journalism are also more commonly associated with male Facebook users.

To validate our findings, we contrasted STEM gender balance estimates from Facebook with data from the 2021 Global Gender Gap Report (WEF, 2021). Our approach enabled the inclusion of 48 countries not covered by the official report.

Among the 152 countries with overlapping data, we observed a correlation between higher offline gender parity and a greater proportion of female users interested in college majors, particularly in non-STEM fields. These results support the viability of using Facebook data as a complementary source for monitoring global gender disparities.

Related Work

According to UNESCO, young women account for only 25% of students in engineering, manufacturing, and construction or information and communication technology in over two-thirds of countries in 2020 (UNESCO, 2020). This STEM gender gap is linked to factors tied to women's self-perception within their social context (Botella et al., 2019; Garcia-Holgado & Garcia-Penalvo, 2022). Initiatives to boost women's recruitment and retention in STEM propose measures to alleviate social identity threats, such as training teachers to encourage STEM vocations in young women and implementing gender-inclusive policies (Garcia-Holgado & Garcia-Penalvo, 2022; Moss-Racusin et al., 2021).

The gender gap in STEM is predominantly examined through surveys (Garcia-Holgado & Garcia-Penalvo, 2022; Tandrayen-Ragoobur & Gokulsing, 2021). For instance, Tandrayen-Ragoobur and Gokulsing (2021) conducted surveys targeting undergraduate students and women working in STEM fields, identifying factors such as family environment, teacher-student relationships, and a sense of community as key influences in shaping career choices. Similarly, Garcia-Holgado and Garcia-Penalvo (2022) developed a model aimed at improving women's attraction to, access to, and retention in STEM within higher education institutions, based on survey data collected in Latin American countries.

Despite their importance, surveys require considerable time and financial resources. As a scalable alternative, researchers have turned to Facebook Ads data to assess diverse demographic characteristics using advertisement audience estimates. This approach has been applied to study lifestyle diseases (Araujo et al., 2017), rural-urban inequalities in difficult-to-reach Italian population groups (Rama et al., 2020), cultural influences on migration across countries (Vieira et al., 2020, 2022), and gender inequality (Al Tamime & Weber, 2022; Kashyap et al., 2020; Vieira & Vasconcelos, 2021). Gender gaps on Facebook serve as proxies for broader gender inequalities (Weber et al., 2018). Garcia et al. (2018) noted that countries with a low Facebook gender gap correlate with increased economic gender equality. Kashyap et al. (2020) found a strong correlation between gender gaps in internet use, low-level digital skills indicators, and data from Facebook and Google Ads.

In the context of measuring gender gaps, Vieira and Vasconcelos (2021) used Facebook data and interests related to college majors to assess the gender disparities in STEM majors in Brazil. Their findings revealed significant variations in the gender gap among different STEM majors, influenced by women's education level and age. Building on this, Al Tamime and Weber (2022) explored the potential of Facebook and Instagram Ads data to model the decline of the gender gap in STEM across different age groups. The study focused on U.S. cities, utilizing APIs filtered by age, gender, and STEM interests. While noting the limitations of social media

advertising data, the study was restricted to a single country and generic interests. To overcome these limitations and account for well-documented gender differences in preferences (Falk & Hermle, 2018), including those observed on Facebook (Cuevas et al., 2021), our study collected data on 142 STEM and non-STEM interests associated with college majors across 198 countries.

Data and Methods

Our methodology builds on prior research by Vieira and Vasconcelos (2021), extending their approach from a national to a global context. We use Facebook Ads data to estimate the gender balance of users interested in STEM and non-STEM college majors across 198 countries.

STEM and non-STEM college majors: STEM refers to majors in Science, Technology, Engineering, and Mathematics. However, definitions of STEM vary across educational, political-social, and personal contexts (Aguilera et al., 2021; Manly et al., 2018). In this study, we adhere to the classification from the National Center for Education Statistics (NCES)[1] to label college majors as STEM or non-STEM. We retrieved a list of 177 majors, categorized into 15 knowledge areas, from the Handshake platform[2]. We then used the Facebook Ads API to obtain audience size estimates for each of these majors.

Selection of countries: We included all 198 countries where Facebook is available and has a sufficiently large user base. Countries where Facebook is restricted[3] or where a given interest had fewer than 1,000 users (in line with Facebook Ads' privacy-mandated thresholds) were excluded.

Facebook Marketing API

The Facebook Marketing API[4] provides estimates of Monthly Active Users (MAU) segmented by demographic attributes, like age, gender, home location for those who stated location in their Facebook profile, and interests (Kosinski et al., 2015). From 177 majors listed by Handshake, we collected data on 193 related interests on Facebook. To refine the dataset, we remove ambiguous or too generic interests (e.g., Music and Photography). We also excluded interests with audiences below 1,000, resulting in a final set of 142 Facebook interests — 66 categorized as STEM and 76 as non-STEM — based on the NCES taxonomy, as shown in Table 1. Users' interests, as inferred or declared on the platform, are used as a proxy for their preferences towards specific fields of study.

[1] https://www.ice.gov/doclib/sevis/pdf/stemList2022.pdf
[2] https://support.joinhandshake.com/hc/en-us/articles/360019970434-List-of-Major-Groups
[3] https://www.Facebook.com/business/help/1155157871341714?id=176276233019487
[4] https://developers.Facebook.com/docs/marketing-apis

Table 1. College majors grouped into STEM and non-STEM.

	College Majors
STEM	Aerospace Engineering, Agriculture, Agronomy, Animal Science, Astronomy, Automation Engineering, Automotive Engineering, Aviation, Biochemistry, Biological Engineering, Biology, Biomedical Engineering, Biotechnology, Botany, Cartography, Cell Biology, Chemistry, Computer Engineering, Computer Programming, Computer Science, Computer Systems Networking, Construction Engineering, Construction Management, Cyber Security, Data Science, Earth Sciences, Ecology, Electrical Engineering, Energy Engineering, Environmental Engineering, Environmental Management, Epidemiology, Food Science, Forensics, Forestry, Genetics, Geography, Geology, Immunology, Industrial Engineering, Information Systems Management, Kinesiology, Landscape Architecture, Management Science, Marine Biology, Materials Science, Mathematics, Mathematics Education, Mechanical Engineering, Microbiology, Molecular Biology, Natural Resource Management, Network Engineering, Neuroscience Nuclear Engineering, Nursery, Oceanography, Physics, Plant Biology, Plant Sciences, Software Design, Soil Science, Statistics, User Experience, Veterinary Sciences, Zoology
Non-STEM	Accounting, Actuarial, Advertising, Agriculture Business, Agriculture Education, American Sign Language, Anthropology, Applied Arts, Architecture, Art History, Business Administration, Business Analytics, Classical Studies, Consulting, Consumer Science, Counseling, Criminal Justice, Criminology, Culinary Arts, Dentistry, Design, Early Childhood Education, Economics, Education Administration, Elementary Education, Emergency Management, Entrepreneurship, Ethics, Ethnic Studies, Exercise Science, Finance, Financial Management, Foreign Languages, Gender Studies, Government, Graphic Design, History, Homeland Security, Hospital Administration, Human Resources, Human Services, Industrial Design, Interior Design, International Business, International Studies, Journalism, Linguistics, Management, Marketing, Media Studies, Medicine, Music Education, Nursing, Nutrition, Occupational Therapy, Operations Management, Pharmacy, Philosophy, Physical Education, Political Science, Product Design, Psychology, Public Administration, Public Health, Public Policy, Public Relations, Religious Studies, Secondary Education, Social Work, Sociology, Special Education, Speech Pathology, Sport Business, Theatre Arts, Theology, Urban Planning

We collected the estimated number of MAU living in each country and interested in each one of the college majors. However, due to the Facebook API's minimum audience threshold, we excluded all college majors with an audience of fewer than 1,000 in a given country. As a result, the number of majors analyzed varies across countries. Figure 1 shows the number of college majors categorized into STEM and non-STEM for the top 50 countries with the highest available number of college majors (i.e., audience greater than 1,000). Notably, the U.S. and India lead in the number of majors and the proportion of STEM interests. Figure 2 shows the proportion of STEM majors within our final dataset of 142 interests (STEM and non-STEM majors) across countries. In most countries, the number of non-STEM college majors is higher than STEM college majors. However, some countries such as Turkmenistan and Yemen show high proportions of non-STEM interests.

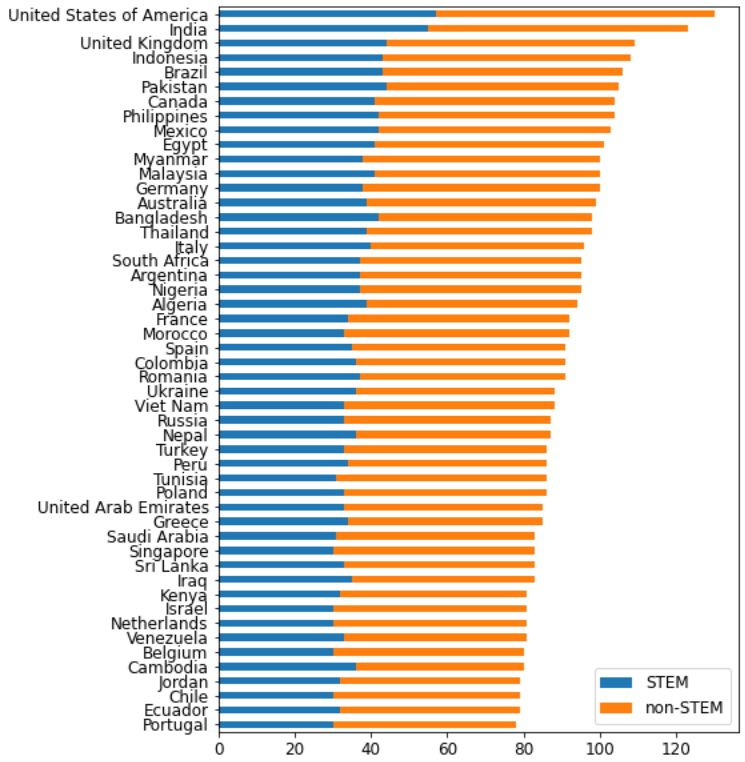


Figure 1. Number of STEM and non-STEM interests available on Facebook per country, depicted for the top 50 countries with over 1,000 MAU associated with college majors.

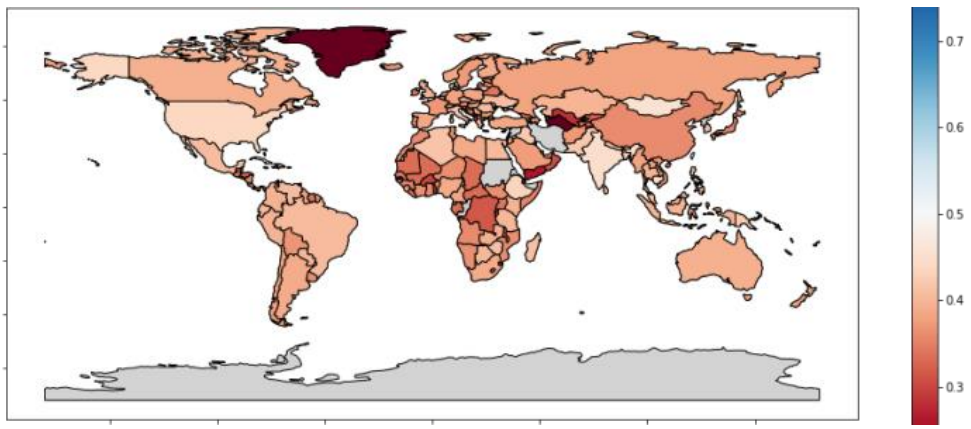


Figure 2. Proportion of STEM majors on Facebook. Colors range from dark red, indicating a higher proportion of non-STEM, to dark blue, indicating a higher proportion of STEM majors. Gray indicates countries with unavailable data.

Gender Balance Metric

To assess the global gender distribution among Facebook users, we compute the **Overall Gender Balance (OGB)**, defined as the proportion of male users in a given population p:

$$OGB_p = \frac{MAU_p(male)}{MAU_p(male) + MAU_p(female)}$$

To assess gender balance in college majors using Facebook users' interests, we adopt the Gender Balance metric proposed in prior studies (Haranko et al., 2018; Vieira & Vasconcelos, 2021). This metric quantifies the ratio of male users interested in a specific major relative to female users. Defining this metric requires specifying the target population. We compute this ratio at the country level and further disaggregate it by major (STEM vs. non-STEM) using additional demographic filters as needed from Facebook Ads. Given a population p, we compute the proportion of users with gender g interested in a college major m as:

$$A_p(g,m) = \frac{MAU_p(g,m)}{MAU_p(g)}$$

Normalization is crucial due to the prevalent imbalanced gender distributions, with more female Facebook users than male, as illustrated in Figure 3a. Subsequently, for the ongoing analysis, we adopt the normalized audience to assess the **Gender Balance (GB)** of a college major m within a population p as:

$$GB_p(m) = \frac{A_p(male,m)}{A_p(male,m) + A_p(female,m)}$$

The GB scores range from 0 to 1, with 0.5 indicating gender parity. Values higher than 0.5 indicate a male majority, while values lower than 0.5 indicate a female majority.

Gender Balance Analysis on Facebook

In this section, we present the OGB and the GB derived from Facebook Ads users' interests in college majors across countries. We start by showing the overall gender balance proportions for each country.

Figure 3a shows OGB values across countries using a color scale that goes from dark red to dark blue. Redder hues denote lower proportions of male users (low OGB values), while bluer shades indicate higher male representation (high OGB values). We use gray to indicate countries with unavailable Facebook data. OGB ranges from 0.39 in Belarus to 0.85 in Yemen (OGB median = 0.51). In most countries, the female audience surpasses the male audience, aligning with prior findings (e.g., Gil-Clavel and Zagheni, 2019) that women are more engaged on the Facebook platform.

Figure 3b presents the median GB values across all majors for each country. GB values range from 0.37 in Georgia to 0.65 in Ethiopia (GB median = 0.48). Notably, 64% of countries exhibit GB scores below 0.5, indicating a higher proportion of female users interested in college majors on Facebook. Exceptions are mainly observed in countries across Africa and Asia. We observe a moderate positive correlation (Pearson's r = 0.45) between the OGB and GB (Figure 5).

Figures 3c and 3d show the median GB values for STEM and non-STEM majors, respectively. A comparison reveals that GB values tend to be higher for STEM than for non-STEM majors, suggesting that in most countries, male users show greater interest in STEM majors. 74% of countries show male-majority interest in STEM, with GB values ranging from 0.37 in New Caledonia to 0.71 in Saudi Arabia (GB STEM median = 0.57). For non-STEM majors, 72% of countries exhibit female-majority interest, with GB values ranging from 0.31 in Georgia to 0.6 in South Sudan (median GB non-STEM = 0.45; 75th percentile = 0.49), reinforcing the trend of greater female interest in non-STEM majors.

Overall, in countries where male users outnumber female users (as shown in Figure 3a), we observe a higher interest in STEM majors among male users (i.e., higher GB for STEM majors), particularly across North Africa and Asia. Only 48 countries have more female users interested in STEM majors (GB STEM < 0.5). For instance, Niger (OGB = 0.8 and GB STEM = 0.39), Tajikistan (OGB = 0.78 and GB STEM = 0.41), Togo (OGB = 0.7 and GB STEM = 0.45), and Yemen (OGB = 0.7 and GB STEM = 0.48) have more male than female Facebook users (i.e., high OGB) and more female than male users interested in STEM.

In contrast, Figure 3d shows that locations where female users are more interested in non-STEM majors also tend to have more male than female users on Facebook. Examples include Tajikistan (OGB = 0.78 and GB non-STEM = 0.38), Azerbaijan (OGB = 0.67 and GB non-STEM = 0.39), Egypt (OGB = 0.63 and GB non-STEM = 0.4), Niger (OGB = 0.8 and GB non-STEM = 0.4), Uzbekistan (OGB = 0.69 and GB non-STEM = 0.41), and Gambia (OGB = 0.66 and GB non-STEM = 0.42). Only 44 countries have median GB values for non-STEM majors higher than 0.5 (i.e., more male than female users are interested in non-STEM majors).

Finally, we find a moderate positive correlation between the two measures, OGB and GB, for both STEM ($r = 0.3$) and non-STEM ($r = 0.45$) majors (see Figure 5). Despite global variation in Facebook usage and the predominantly female-skewed user base (i.e., OGB < 0.5), our observations consistently show higher interest in STEM majors among male users (GB STEM > 0.5) and greater interest in non-STEM majors among female users (GB non-STEM < 0.5).

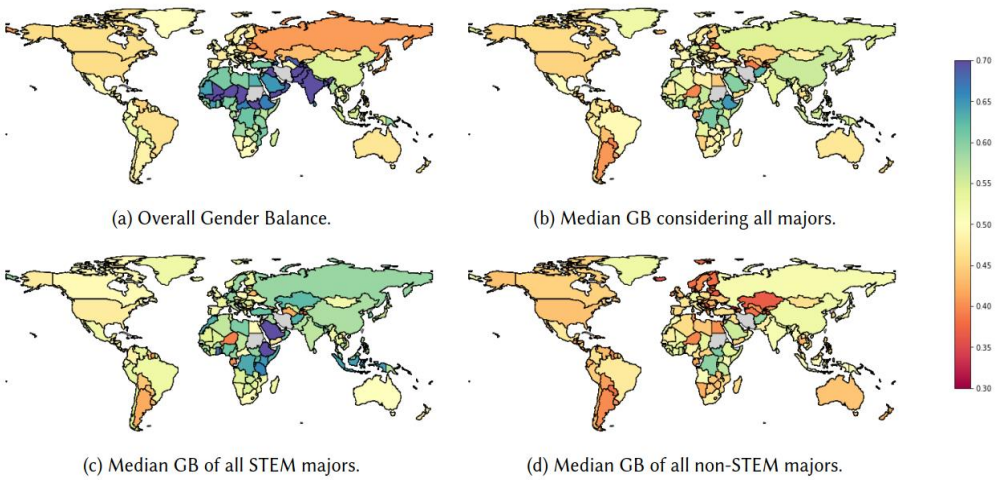


Figure 3. Overall Gender Balance (OGB) and Gender Balance (GB) across countries. Colors range from red, indicating a higher proportion of female users, to blue, indicating a higher proportion of male users. Gray indicates countries with unavailable data.

To provide more detail on the GB values for each major, we selected the top five countries with the highest number of majors (see Figure 1). Figure 4 displays GB values for each major, focusing on Facebook users from these five countries. Figures 4a and 4b illustrate STEM and non-STEM majors, respectively, using the same color scale as in Figure 3—redder and bluer shades represent a greater proportion of female and male users, respectively. White areas indicate majors in specific countries with insufficient Facebook data (i.e., a Facebook audience size of 1,000 users).

In both Figures 4a and 4b, around 60% of the cells show GB > 0.5 for STEM and GB < 0.5 for non-STEM, emphasizing that male users dominate interest in STEM majors, while female users are more interested in non-STEM majors. However, some STEM majors (e.g., Life Sciences and Mathematics) attract more female than male users.

Consistent with findings by Vieira and Vasconcelos (2021), we observe two distinct patterns across STEM majors: (i) Engineering and Technology interests are predominantly male-dominated (i.e., high GB), while (ii) Science and Math majors exhibit a high number of female users (i.e., low GB). This highlights that despite the overall gender gap in STEM, the pattern does not apply uniformly across all STEM majors. Therefore, when designing policies or interventions aimed at increasing female participation in STEM, it is essential to consider the variability in gender balance across different majors. Lastly, even within non-STEM majors, there are exceptions where male users outnumber female users (i.e., high GB), such as in Economics and Business, History, Government, and Journalism.

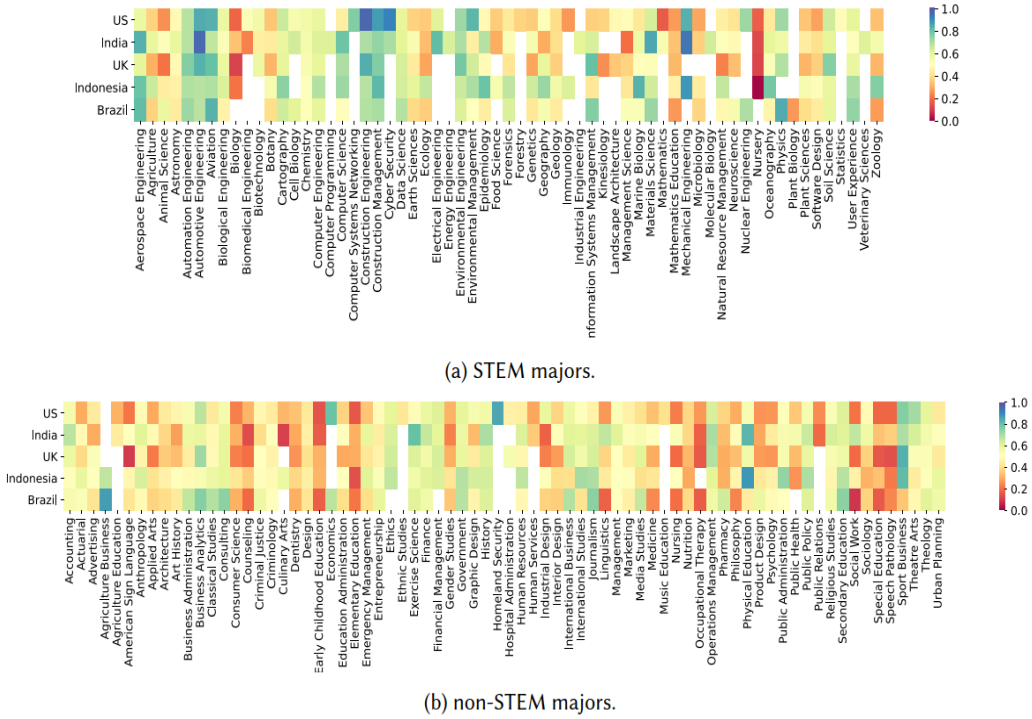


Figure 4. Gender Balance (GB) for each major in the top five countries with the highest number of majors. Colors range from red (lower GB) to blue (higher GB). White indicates unavailable data.

Contrasting online and offline gender gaps

The contrast between online and offline gender gaps can offer valuable insights into the interconnectedness of online and offline measures of gender gap and shed light on the effectiveness of using social media data to measure gender gaps. Offline indicators can provide a benchmark for assessing the extent to which social media data can capture the gender gap while also highlighting any methodological limitations.

To facilitate this comparison, we used data from the 2021 report provided by the World Economic Forum (WEF, 2021). The Global Gender Gap Index (GGGI) covers 156 countries and comprises four sub-indices: Economic Participation and Opportunity, Educational Attainment, Health and Survival, and Political Empowerment. GGGI values range from zero (complete disparity) to 1 (complete parity). Figure 5 presents Pearson's correlation coefficients between the gender balance indicators derived from Facebook data and the GGGI (including its sub-indices) for the 152 countries available in both datasets.

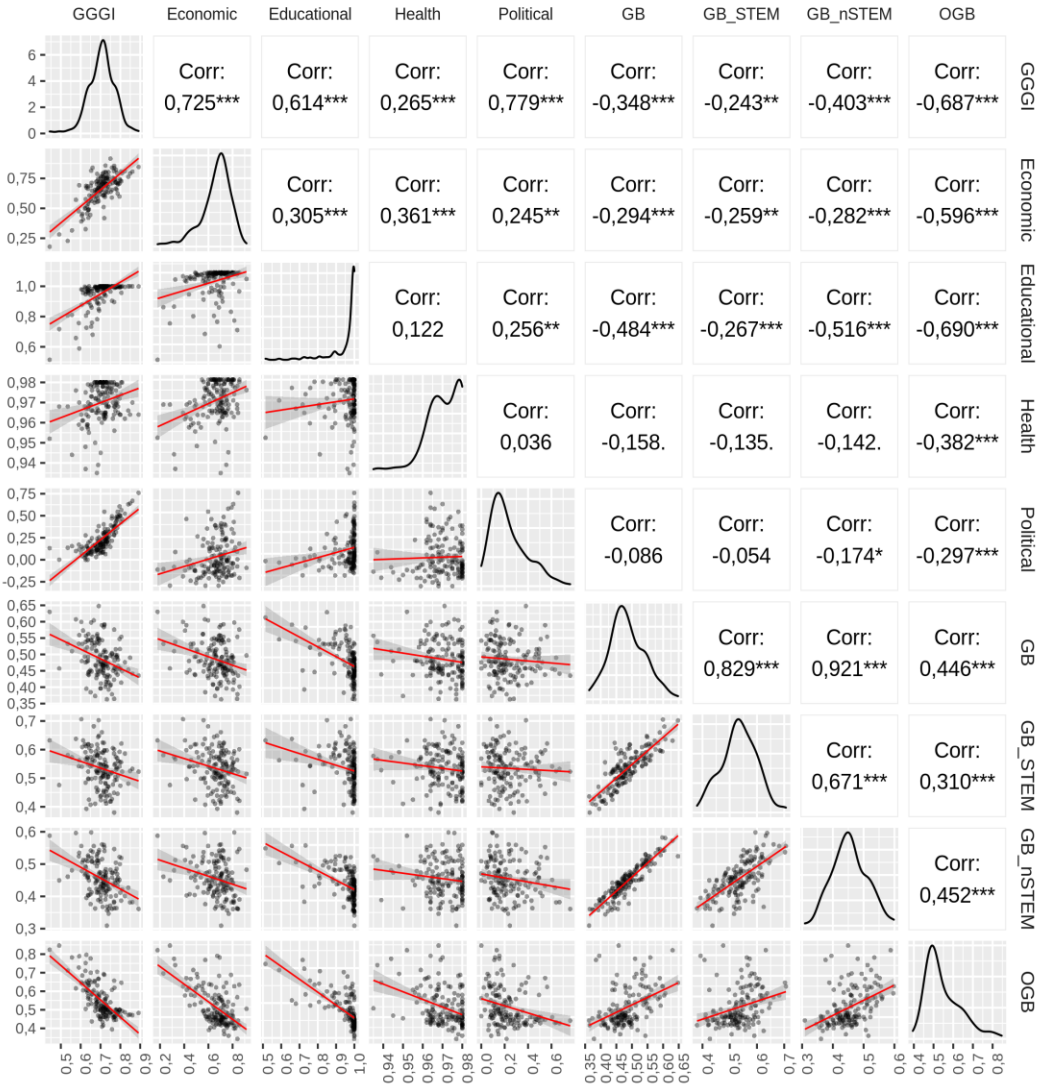


Figure 5. Correlations among Gender Balance (GB) measures based on Facebook data, the Global Gender Gap Index (GGGI), and its components: Economic Participation and Opportunity, Educational Attainment, Health and Survival, and Political Empowerment.
∗ ∗ ∗p < 0.001; ∗ ∗ p < 0.01; ∗p < 0.05

Facebook OGB and the GGGI show a strong negative correlation (r = −0.69). Countries with high GGGI values, such as Iceland, Finland, Norway, New Zealand, and Sweden, approach gender parity and have a higher proportion of female than male Facebook users (i.e., lower OGB). In contrast, countries with low GGGI values, such as Yemen and Afghanistan, have more male than female Facebook users. However, exceptions exist. Countries like Bangladesh, the United Arab Emirates, Burundi, Rwanda, Albania, and Mozambique demonstrate high levels of gender parity offline (high GGGI) but still have predominantly male Facebook users (i.e.,

high OGB), pointing to potential limitations in digital access or platform-specific dynamics.

We also find a moderate negative correlation between Facebook GB and the GGGI ($r = -0.35$), indicating that in countries with greater gender disparity (low GGGI), men are more likely to express interest in higher education majors online. The correlation is slightly stronger when focusing on non-STEM majors ($r = -0.4$), suggesting that in more gender-equal countries, female users show greater interest in non-STEM majors on Facebook.

As previously noted, countries with high GGGI values (e.g., Iceland, Finland, Norway, New Zealand, and Sweden) also exhibit some of the lowest GB values overall—especially in non-STEM majors—implying a larger presence of female users expressing academic interests. In contrast, Afghanistan stands out as a country with both low gender equality and a high proportion of male users interested in college majors (high GB). Yemen, however, represents an anomaly: despite a low GGGI, it has more female than male Facebook users interested in majors (i.e., low GB).

Finally, we found a low correlation between the GGGI and STEM GB values ($r = -0.24$). This may be due to the aggregation of STEM majors: high GB values in Engineering and Technology are counterbalanced by lower values in Life Sciences and Mathematics (see Figure 4). To capture these divergent patterns more accurately, future work should treat STEM as two distinct categories.

Discussion and Conclusion

This study uses Facebook Ads data to assess the overall gender balance and distribution of interest in STEM majors across 198 countries. Our findings confirm a general female user bias in the Facebook audience, consistent with previous research, with notable exceptions in specific African and Asian countries. While most countries exhibit a higher proportion of female users expressing interest in various academic majors, STEM majors show a predominance of male users in 74% of countries, in contrast to non-STEM majors, where female users dominate in 72% of cases. Within STEM, gender patterns vary substantially: Life Sciences and Mathematics attract more female interest, whereas Engineering and Technology remain male-dominated. Similarly, some non-STEM majors—such as Economics and Business, History, Government, and Journalism—tend to be more popular among male users.

Our study introduces a timely, cost-effective, reproducible, and scalable methodology to assess global gender disparities in STEM using digital trace data. These insights offer important implications for policymakers, industry leaders, and educational institutions aiming to address gender inequality. In future work, we plan to validate these findings against offline gender gap indicators and expand the scope to other social media platforms by utilizing their APIs to capture user-level attributes like gender, education, and interests.

Our analysis relies exclusively on publicly available data from the Facebook Marketing API, adhering to ethical guidelines (Rivers & Lewis, 2014). The data is aggregated and anonymized, ensuring compliance with Facebook's terms of

service[5]. While this approach demonstrates the feasibility of using Facebook Ads data for demographic research, several limitations remain. First, interests are either self-declared by users or inferred by Facebook based on behavioral signals (e.g., posts, likes, shares), and the exact inference mechanisms remain unclear. Despite this, we assume that users' interests in college majors on Facebook are a good proxy for studying the gender gap across disciplines.

Second, demographic attributes are restricted to those offered by the Facebook Ads Platform, treating gender as a binary variable. Third, our analysis focused on country-level gender balance without an age group breakdown. We aimed to maximize data coverage by avoiding requests that excessively narrow the Facebook audience. When the estimated audience size falls below Facebook's threshold of 1,000 users, the actual number could range from 0 to 1,000, introducing uncertainty. Fourth, cultural and contextual factors likely shape Facebook usage patterns, which may affect our results. Moreover, the classification of what constitutes a STEM field is itself contested. Definitions vary across stakeholders—such as educators, policymakers, and industry representatives—as well as by context, including immigration policies targeting STEM workers. For this reason, we designed a flexible and reproducible methodology that can accommodate different STEM classification schemes. We believe our results are largely robust to these definitional differences.

While platforms like LinkedIn might also offer valuable data for measuring gender gaps, they primarily reflect labor market participants (Najafikhah & Shamizanjani, 2018), which is outside the scope of this study. Facebook, by contrast, remains the world's largest social media platform—with particularly strong usage among young people (Duggan, Brenner, et al., 2013)—allowing us to capture a broader spectrum of users, including those not currently enrolled in or employed in STEM fields. By focusing on users' expressed interests rather than occupational or educational status, this study offers a complementary perspective on gender gaps in STEM interest.

References

Aguilera, D., Lupiáñez, J., Vílchez- González, J., & Perales-Palacios, F. (2021). In search of a long-awaited consensus on disciplinary integration in STEM education. Mathematics, 9 (6), 597.

Al Tamime, R., & Weber, I. (2022). Using social media advertisement data to monitor the gender gap in STEM: opportunities and challenges. Peer J Computer Science, 8 (994).

Alexander, M., Polimis, K., & Zagheni, E. (2019). The Impact of Hurricane Maria on Out-migration from Puerto Rico: Evidence from Facebook Data. Population and Development Review, 45 (3), 617–630.

Araujo, M., Mejova, Y., Weber, I., & Benevenuto, F. (2017). Using Facebook ads audiences for global lifestyle disease surveillance: Promises and limitations. WebSci'17.

Botella, C., Rueda, S., López-Iñesta, E., & Marzal, P. (2019). Gender diversity in STEM disciplines: A multiple factor problem. Entropy, 21, 30.

Cheryan, S., Ziegler, S., Montoya, A., & Jiang, L. (2016). Why are some STEM fields more gender balanced than others? Psychological Bulletin, 143.

[5] https://developers.Facebook.com/policy/\#marketingapi

Cuevas, A., Cuevas, R., Desmet, K., & Ortuño-Ortín, I. (2021). The gender gap in preferences: Evidence from 45,397 Facebook interests (tech. rep.). National Bureau of Economic Research.

Duggan, M., Brenner, J., et al. (2013). The demographics of social media users, 2012 (Vol. 14). Pew Research Center's Internet & American Life Project Washington, DC.

Falk, A., & Hermle, J. (2018). Relationship of gender differences in preferences to economic development and gender equality. Science, 362 (6412). Forbes. (2018). A study finds that diverse companies produce 19% more revenue.

Garcia, D., Mitike K., Y., Cuevas, A., Cebrian, M., Moro, E., Rahwan, I., & Cuevas, R. (2018). Analyzing gender inequality through large-scale Facebook advertising data. 115 (27), 6958–6963.

Garcia-Holgado, A., & Garcia-Penalvo, F. (2022). A model for bridging the gender gap in STEM in higher education institutions. In Women in STEM in higher education: Good practices of attraction, access and retainment in higher education (pp. 1–19).

Gil-Clavel, S., & Zagheni, E. (2019). Demographic differentials in Facebook usage around the world. ICWSM'19.

Gompers, P., & Kovvali, S. (2018). The other diversity dividend. Harvard Business Review, 96 (4), 72–77.

Guimarães, S., Reis, J., Vasconcelos, M., & Benevenuto, F. (2021). Characterizing political bias and comments associated with news on Brazilian Facebook. Social Network Analysis and Mining, 11 (1), 94.

Haranko, K., Zagheni, E., Garimella, K., & Weber, I. (2018). Professional gender gaps across us cities.

Kashyap, R., Fatehkia, M., Al Tamime, R., & Weber, I. (2020). Monitoring global digital gender inequality using the online populations of Facebook and Google. Demographic Research, 43 (27), 779–816.

Kosinski, M., Matz, S., Gosling, S., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. American Psychologist, 70 (6), 543.

Manly, C., Wells, R., & Kommers, S. (2018). The influence of STEM definitions for research on women's college attainment. International Journal of STEM Education, 5 (1), 45.

Mejova, Y., Gandhi, H., Rafaliya, T., Sitapara, M., Kashyap, R., & Weber, I. (2018). Measuring subnational digital gender inequality in India through gender gaps in Facebook use. SIGCAS'18.

Moss-Racusin, C., Pietri, E., Toorn, J., & Ashburn-Nardo, L. (2021). Boosting the sustainable representation of women in STEM with evidence-based policy initiatives. Policy Insights from the Behavioral and Brain Sciences, 8 (1), 50–58.

Munoz-Boudet, A. M., & Revenga, A. (2017). Breaking the STEM ceiling for girls.

Najafikhah, S., & Shamizanjani, M. (2018). Examining the motivations of LinkedIn users and their demographics. 5th European Conference on Social Media ECSM 2018, 171.

Palotti, J., Adler, N., Morales-Guzman, A., Villaveces, J., Sekara, V., Herranz, M., Al-Asad, M., & Weber, I. (2020). Monitoring of the Venezuelan exodus through Facebook's advertising platform. PLOS ONE, 15 (2).

Rama, D., Mejova, Y., Tizzoni, M., Kalimeri, K., & Weber, I. (2020). Facebook ads as a demographic tool to measure the urban-rural divide. TheWebConf'20.

Rivers, C. M., & Lewis, B. L. (2014). Ethical research standards in a world of big data. F1000Research, 3 (38), 38.

Tandrayen-Ragoobur, V., & Gokulsing, D. (2021). Gender gap in STEM education and career choices: What matters? Journal of Applied Research in Higher Education.
UNESCO. (2020). Global education monitoring report: Gender report, a new generation: 25 years of efforts for gender equality in education [Accessed: 2022-08-08].
UNESCO. (2021). Women a minority in industry 4.0 fields [Accessed: 2022-08-08].
Vieira, C., Lohmann, S., Zagheni, E., Vaz de Melo, P. O. S., Benevenuto, F., & Ribeiro, F. N. (2022). The interplay of migration and cultural similarity between countries: Evidence from Facebook data on food and drink interests. PLOS ONE, 17 (2), 1–21.
Vieira, C., Ribeiro, F., de Melo, P., Benevenuto, F., & Zagheni, E. (2020). Using Facebook data to measure cultural distance between countries: The case of Brazilian cuisine. TheWebConf'20.
Vieira, C., & Vasconcelos, M. (2021). Using Facebook ads data to assess gender balance in STEM: Evidence from Brazil. TheWebConf'21.
Weber, I., Kashyap, R., & Zagheni, E. (2018). Using advertising audience estimates to improve global development statistics. Itu Journal: Ict Discoveries, 1 (2).
WEF. (2016). Women and work in the fourth industrial revolution [Accessed: 2022-08-08].
WEF. (2021). Global gender gap report [Accessed: 2022-09-14].

https://doi.org/10.51408/issi2025_163

Citation Context Analysis: Evaluating Human vs. AI Annotations in Gameplay Bricks Research

Marc Bertin[1], Julian Alvarez[2], Thierry Lafouge[3]

[1]*marc.bertin@univ-lyon1.fr*
ELICO, Université Lyon1, Lyon, Villeurbanne (France)

[2]*julian.alvarez@univ-Lille.fr*
GERiiCO, Université de Lille, Lille (France)
Immersive Factory, R&D, Paris (France)

[3]*thierry.lafouge@univ-lyon1.fr*
ELICO, Université Lyon1, Lyon, Villeurbanne (France)

Abstract

Recent advances in citation analysis have moved beyond traditional bibliometric approaches to explore the contextual roles of citations in academic discourse. While Large Language Models (LLMs) offer new possibilities for analyzing citation contexts, challenges persist regarding annotated dataset availability and inherent biases in citation categorization schemes. This study presents a novel comparative analysis of citation contexts, focusing on the gameplay bricks framework developed by Alvarez and Djaouti (2006) across a ten-year period (2008-2018).

Our research employs prompt engineering techniques to analyze citation contexts in nine languages, comparing human expert annotations with ChatGPT-generated analyses. This micro-level investigation examines how the gameplay bricks model has been referenced, appropriated, and critiqued across different disciplines. The study addresses three primary research questions: the interpretation of citation contexts by domain experts, the alignment between AI-generated categorizations and expert judgments, and the insights gained from comparing human and AI annotations in multilingual scientific discourse.

The methodology combines traditional human annotation with AI-assisted classification through prompt-based methods. Our analysis reveals a predominance of definition and appropriation categories, indicating widespread adoption of the gameplay bricks model across disciplines. Computer science publications showed higher rates of model appropriation, while humanities disciplines demonstrated more critical engagement. The study identified particular challenges in capturing neutrality and criticism, attributable to both AI model limitations and the inherent complexity of citation context interpretation.

Results demonstrate that while ChatGPT-powered annotation offers scalability advantages, it faces limitations in processing contextual nuances and interpretive depth, particularly across different languages. The comparative analysis highlights discrepancies between human and AI interpretations, suggesting the need for hybrid approaches that leverage both human expertise and AI capabilities. These findings contribute to ongoing discussions about AI's role in academic discourse analysis and raise important questions about citation practices.

It provides insights into the evolution of academic discourse around the gameplay bricks framework while highlighting methodological considerations for future citation analysis studies. The findings underscore the importance of developing more sophisticated tools for citation context analysis that can account for linguistic and disciplinary variations.

Introduction

Since a comprehensive theory of citation has not yet been established, emerging models are being proposed, as highlighted by Tahamtan & Bornmann (2022) with *The Social Systems Citation Theory (SSCT): A proposal to use the social systems theory for conceptualizing publications and their citation links*. Research on citation contexts focuses on two interrelated and complementary aspects: a conceptual dimension that leads to the proposal of schemas, categories, and functions related to the nature of citation acts, and an operational dimension of these categories, which requires the implementation of corpora, computational modeling, and annotation evaluation.

Today, numerous citation categories have been proposed and many surveys written such as Bornmann & Daniel 2008, Hernández-Alvarez & Gomez 2016, Zhang et al 2023 to name but a few, and computational tools are becoming more powerful with the advent of large language models (LLMs). This paradigm shifts in methods derived from natural language processing (NLP) opens new perspectives. However, two main limitations remain: the scarcity of annotated resources for training machine learning methods and the nature of the categories on which supervised approaches rely. For this article, drawing on recent advancements in prompt engineering, we propose a case study to explore the relevance of analyzing citation contexts over a ten-year period for a specific research topic: the *gameplay bricks* framework introduced by Julian Alvarez and Damien Djaouti in 2006. Our study focuses exclusively on the analysis of citation contexts from a temporal perspective. Understanding citation contexts requires examining their evolution over time. At a micro-level—opposed to the macro and meso approaches of traditional bibliometric studies—returning to the text and conducting a fine-grained analysis are essential for understanding controversies and debates. For example, a dispute between two researchers through successive articles can only be analyzed by reading the texts in full, even if the citation frequency is low. A highly cited article, on the other hand, indicates high visibility, meaning it appears in many bibliographies. But what role does it play for the citing researcher? Why and where was it cited? Studies have shown that the rhetorical structure of a scientific article significantly influences the nature of citation contexts, depending on whether the citation appears in the introduction (literature review), methodology, results, or discussion sections.

The study we propose confronts human expertise—represented by an identified researcher in the field of gameplay studies—with the latest AI approaches using prompts, specifically ChatGPT, to analyze citation contexts within this corpus. To illustrate our approach, we selected a case study based on the *gameplay bricks* framework introduced by Julian Alvarez and Damien Djaouti in 2006. The advantage of this work is that it includes an inventory of international citations, which have been analyzed by one of the researchers to identify citation contexts. The goal is to determine whether these citations reflect an adoption of the *gameplay bricks* model, a critique of it, or a neutral stance (Alvarez, 2018). This provides a basis for a comparative study between human expertise and AI analysis. Additionally, this corpus offers other advantages, such as its multilingual nature, making it suitable for a first iteration of a comparative study between human and AI analyses.

Our approach thus leads us to examine the nature of citations received by various research articles. For example, we may explore the fundamental nature of citations related to gameplay studies: Are they negative, positive, or neutral? Can we identify cases of conceptual appropriation, and if so, of what kind? Which scientific fields refer to the studied works? Are these works cited in other languages? In the latter case, what functions do the citation contexts convey?
We have therefore chosen to compare the analysis performed by a researcher with that of an AI on the same corpus to better understand the identified and generated citation contexts in both cases. Beyond the question of reliability, we also consider it relevant to leverage such a comparative analysis to uncover the insights that such a cross-analysis can provide. This constitutes the primary objective of this article. In the context of using AI to help identify citation contexts, the underlying issue of reliability will be addressed in the evaluation section.

GPT and LLM Litterature review

Over the past few decades, automatic classification of citation features has evolved in parallel with advances in natural language processing (NLP) technologies. However, despite numerous studies documented in surveys (Bornmann and Daniel (2008); Hernández-Alvarez and Gomez (2016); Jha, Jbara, Qazvinian, and Radev (2017); Lyu, Ruan, Xie, and Cheng (2021), significant limitations and persistent biases hinder its widespread adoption. One of the most noteworthy advancements in natural language processing (NLP) is the emergence of large language models, such as GPT (Generative Pre-trained Transformer) (Radford (2018), 2019; Brown et al. (2020a)), along with their numerous iterations. These models have exhibited exceptional capabilities across a diverse range of linguistic tasks. Typically, GPT models undergo two key phases: pre-training on extensive text datasets to learn general language patterns, followed by fine-tuning for specific downstream tasks to generate highly human-like language. Among the various fine-tuning techniques, prompt engineering stands out as a particularly accessible approach for non-specialists, offering a user-friendly means to harness the potential of these powerful models. While Nishikawa's research demonstrated the consistency of LLMs in this task, it also highlighted the limits of their ability to fully replace human annotators. Indeed, Lahiri et al. (2023) introduce CitePrompt, a novel tool leveraging prompt learning for citation intent classification. By optimizing the choice of pretrained language models, prompt templates, and verbalizers, CitePrompt achieves state-of-theart performance on the ACL-ARC dataset and significant improvements on SciCite, requiring minimal external document information. They propose a first-of-its-kind approach to adapt citation intent classification to few-shot and zero-shot settings, addressing the scarcity of large labeled datasets.

Zero-based and low-based learning for labeling citation contexts

Nevertheless, emerging approaches such as Zero-Shot and Few-Shot Learning for citation labeling, inspired by the work of Brown et al. (2020b), offer promising avenues for exploration. In fact, the literature shows that other fields use this type of approach to compensate for the lack of annotated corpora. ChatGPT offers a broad

spectrum of applications in the field of research, particularly in the domain of text mining. For example, Mathebula, Modupe, and Marivate (2024) in sentiment analysis for financial applications, enhancing the accuracy and utility of customer feedback in shaping business decisions. Khan, Khan, Li, Ullah, and Zhao (2025) introduces a novel approach using ChatGPT as both annotator and negotiator, achieving a 94% accuracy rate with deep learning classifiers in detecting emotions in negative reviews from low-rated apps, demonstrating the potential of generative AI in enhancing annotation reliability and performance. Chen et al. (2023) evaluates ChatGPT's performance on biomedical tasks through a comprehensive benchmark involving article abstracts, clinical trial descriptions, and biomedical questions, demonstrating its effectiveness and versatility in biomedical text comprehension, reasoning, and generation. Zhu et al. (2022) have described the fundamental concepts underlying this approach, which could play a central role in advancing citation analysis. More recently, Lahiri, Sanyal, and Mukherjee (2023) have positioned Prompt Learning as a particularly suitable method for tackling this challenge.

Discussion of Gameplay Bricks Model

The distinction between the concepts of video games and Serious Games is based on principles modeled by Alvarez, Djaouti, Ghassempouri, Jessel, and Methel (2006). This model, named "Gameplay Bricks", was originally designed to deconstruct video games in an effort to both classify video games and identify characteristics that could distinguish Serious Games from video games within a formal system Alvarez et al. (2006). After 2006, the Game Bricks model was consolidated over the period from 2007 to 2010. The literature on which this model is based is presented in Table 2. While the core of the model will be repeated in the literature, it is interesting to note the variety of media used to build the Game Bricks model. More than a decade after its introduction into the scientific community, how has the Gameplay Bricks model been perceived, used or criticized? What specific criticisms can we identify from the citations collected? The corpus for this study will come to an end in 2018, when a synthesis book will be published on this issue. In 2024, we will have the necessary hindsight and coverage to observe the spread of this model within the various scientific communities. Indeed, the choice of this model is even more interesting in that it is mobilized through numerous national and international citations, in different languages and in different contexts. The game bricks expert was able to build up a categorization of citation functions through manual study and human expertise.

Problems

Furthermore, the underlying question that interests us in this study is whether, given the current state of research on citation contexts, we are capable of producing semantic annotations of citation contexts that would ultimately allow us to track the dissemination of models, as demonstrated in this study, or theories, as well as the identification of controversies. Our research problem is as follows: Based on the corpora generated around the modeling of Game Bricks, can we analyze citation contexts and derive categories that align with expert-produced knowledge?

As highlighted in the literature review, we face two major limitations: the lack of stable categorizations and a bias introduced by supervised approaches, which still lack annotated corpora covering all categories and disciplines. Recent state-of-the-art reports shed light on the latest studies in this field. The Gameplay Bricks model is employed to determine whether ChatGPT-based approaches applied to citation contexts can provide an application framework for understanding discussions, or even controversies, surrounding this model. To assess ChatGPT's potential in research and its application to citation contexts, our study explores its understanding of semantic usage, focusing on specialized topics related to gameplay bricks. Based on these elements, we propose the following three main research questions:

1. How does a domain expert mobilize citation contexts?
2. Do the categorizations produced by ChatGPT agree with the expert?
3. How to navigate in a contextualized space of citation contexts?

The content of this paper is organized as follows: Section 1 provides a general introduction, including a state-of-the-art review and the research problem addressed. Section 2 describes the dataset constructed around Gameplay Bricks and outlines our experimental approach to citation context categorization, which integrates human expertise with the proposed solution using OpenAI ChatGPT. Section 3 presents the results obtained from AI-generated citation context annotations, the resulting graph, and a human analysis focusing on cases of appropriation. Section 4 offers a discussion comparing the human analysis with the proposed OpenAI ChatGPT approach.

Methodologies

As we have just observed, the potential of this type of approach is evident. Two studies that have particularly drawn our attention in constructing our methodology are the studies of Lahiri, A., Sanyal, D.K., Mukherjee, I. (2023) and the latest research from Nishikawa, K., & Koshiba, H. (2024), which explores the application of large language models (LLMs) to citation context analysis. The article of Nishikawa, K., & Koshiba, H. (2024) highlights a crucial limitation of current approaches: the lack of annotated corpora. The study emphasizes the experiment's inability to achieve relevant annotation results. This research employed five classes for *Citation Purpose* when citing a referenced paper: Background, Comparison, Critique, Evidence, and Use. Regarding *Citation Sentiment*—which refers to the mental attitude of the author of a citing paper towards the cited paper—the authors used three classes: Positive, Negative, and Neutral. The choice of categories is based on the availability of an annotated corpus and resources for the scientific community.

Designing Prompts for Citation Context Classification

One of the key aspects of this approach with Large Language Models (LLMs) is the design and application of prompts, which are structured natural language inputs that guide the model's response. In this context, the structure and specificity of the prompt significantly influence the quality and relevance of the generated output. To ensure

optimal performance, it is essential to provide input categories that do not involve intrinsic complexity or rely on excessively broad generalizations. The prompt must explicitly instruct the LLM to focus on classification, ensuring that the model produces the expected results. To further enhance prompt precision, several strategies can be employed. One effective approach is to provide explicit examples of citation contexts. Instead of relying on a single query, using labeled citation contexts allows the model to generalize more effectively and improve the relevance of its outputs. Still in the context of the implementation of our method, we have taken on board the remarks of Nishikawa 2024 concerning techniques for improving results, namely the few-shot Brown et al. (2020) or chain-of-thought approaches of Wei et al. (2022) and Zhang (2023) fot Chain-of-thought with ChatGPT for Stance Detection on Social Media.

Nishiwaka's methodological approach is structured around four citation incentive models that build upon the basic instructions by progressively incorporating additional contextual elements. The first model includes only class types, providing a minimal framework for classification. The second model expands on this by integrating class types along with their definitions, offering greater conceptual clarity. The third model further refines the approach by including annotation procedures, ensuring more precise and standardized applications. The most comprehensive model incorporates class types, definitions, annotation procedures, keywords, and example sentences, creating a fully detailed framework for citation analysis. This final model closely resembles the manual used in their previous study (Nishikawa, 2023).

In line with these results, it is essential to develop a prompt-based approach that takes these constraints into account. To achieve this, we need to produce a classification for citation contexts that provides a clear and structured framework for annotation. This leads to the next point, which is to identify the categories that can meet the prompting requirement.

Construction Citation Context Classification

Previous studies provide a certain richness despite inherent biases in their construction and design, such as corpus size, disciplinary scope, or unaddressed biases, such as the language used. We can cite the work of Teufel (2006), Athar (2011), Dong & Schäfer (2011), Bertin & Atanassova (2024) as key references for this study. Our approach focuses on the various categories proposed in the literature to design prompts that provide classifications aligned with the discursive forms likely to appear in citation contexts. These classifications aim to minimize ambiguity and abstraction while accurately reflecting the nature of citations. To achieve this, we draw on the work of Liu et al. (2023), and its applications in identifying citation intents in scientific papers, as explored in the recent study (Nishikawa, K., & Koshiba, H., 2024).

The categories used to build the chatGPT prompt naturally draw on the labels proposed by the expert, but also on other categories identified in the literature. The categories selected must convey notions that can be identified in a citation context. To this end, we have selected categories identifiable by discourse forms likely to be

present in citation contexts. Based on established classifications, we have produced a selective and descriptive list of functions. These functions are likely to be relevant to the processed corpus and implementable via prompts for the intended task. We have synthesized the main citation functions, as outlined in the following table 1:

Table 1. A Synthesis of Citation Functions: Categories and Their Discursive Roles.

Category	Description
Definition	The citing paper provides a definition of a concept from the cited paper.
Method	A citation instance where the cited work provides the method or technique used in the citing paper, which either describes or applies the methodology introduced in the cited work.
Hypothesis	The cited work is used to support or inspire a hypothesis in the citing work.
Extension	The citing work's research work is an improvement or extension of the cited work.
Comparison	A citation instance that involves any form of comparison or contrast between different cited papers or between the cited work and the citing paper. It highlights similarities or differences between the cited work and the author's own research.
Agreement	The citing paper explicitly agrees with or endorses the cited paper's conclusions.
Result	A citation instance in which the citing work mentions specific results or general findings of the cited paper.
Extension	The citing paper extends the methods, tools, or data of the cited paper.
Point of view	The cited work is used to illustrate a particular theoretical or conceptual perspective
Future	The cited paper may be a potential reference for future work.

The Gameplay Bricks Corpus

The distinction between video games and Serious Games is based on principles formalized by Alvarez et al. (2006) through the "Gameplay Bricks" model. Initially developed to deconstruct video games, this model aimed to establish a classification system while identifying specific characteristics that differentiate Serious Games from traditional video games (see Alvarez et al., 2006). Following its introduction, the Gameplay Bricks model was further refined between 2007 and 2010. Table 2 provides a synthesis of the literature that contributed to the development of this model. While its foundational principles are consistently referenced in subsequent research, it is particularly noteworthy that a diverse range of media has been utilized in shaping and expanding the Gameplay Bricks framework.

Table 2. Works published between 2006-2010.

Categories	References
Conference	Alvarez J., D. Djaouti, and R. Ghassempouri (2006), "*Morphological study of videogames*," CGIE'06 conference, Australia.", 2006
Conference Proceedings	Djaouti, Damien, J Alvarez, Jp Jessel, Gilles Methel, and P Molinier. 2007. "*The Nature of Gameplay: A Videogame Classification*." Cybergames Conference, no. July 2015.", 2007
Conference Proceedings	Djaouti, D., Alvarez, J., Jessel, J.-P., & Methel, G. (2007). *Towards a classification of video games*. Artificial and Ambient Intelligence convention (Artificial Societies for Ambient Intelligence) (AISB (ASAMi) 2007).", 2007
Conference Proceedings	Alvarez et al., 2007] Alvarez, J., Djaouti, D., Jessel, J.-P., Methel, G. et Molinier, P. (2007). *Morphologie des jeux vidéo*. In H2PTM, Hammamet, Tunisie, 29/10/2007-31/10/2007, numéro 978-2-7462-1891-8 de Lavoisier, pages 277–287, http://www.editions-hermes.fr/. Hermès Science Publications.", 2007
Thesis	Thesis,"Alvarez, J. (2007). *Du jeu vidéo au serious game, approches culturelle, pragmatique et formelle*, Thèse de doctorat en science de l'information et de la communication, Toulouse, France : Université de Toulouse.", 2007
Article	Djaouti, D., Alvarez, J., Jessel, J.-P., and Methel, G. (2008). *Play, Game, World: Anatomy of a Video-Game*. International Journal of Intelligent Games & Simulation, 5(1):35–36.", 2008
Book	Book,"Alvarez, J., & Djaouti, D. (2010), "Introduction au Serious Game", Questions théoriques, vol. 1, Paris.", 2010
Website	Website,"Alvarez, Julian et Damien Djaouti. S.d. Game Classification: la classification en ligne du jeu vidéo. <http://www.gameclassification.com/>.", 2010

The first step was to create an average from the WoS based on the search equation to build a corpus: Gameplay Bricks (All Fields) and CMN-3138-2022 (Auhtor Identifiers) or AAE-9793-2019 (Author identifiers) which produced 6 references for 45 citations from 2007 to 2023. For this study, which covers the period from 2008 to 2018, the Web of Science (WoS) database reports 3 articles with a total of 19 citations. From this equation, the results were extended via other databases to cover the multilingual aspect. As the concept is mobilized by the international community and has a coverage that goes beyond English-language publications, it was important to extend our research to have a consolidated corpus. We identified a total of 47 scientific articles in 9 languages and 40 theses in 4 languages, highlighting the richness and international scope of the concept explored in this study, using additional resources, databases, as well as laboratory and institutional websites. Using Google Scholar with keywords such as Brique Gameplay and Gameplay Brick combined with the names Djaouti or Alvarez, nearly 200 national and international references were identified in 2018. Among these results, self-citations were removed, ensuring that the same author was cited only once, with preference given

to the oldest or most detailed article referring to the Gameplay Brick model. In addition, articles mentioning the notion of bricks without referring to or using the model were excluded. Indeed, the term brick is often used in everyday language to refer to the idea of a component. This process resulted in a final count of 47 articles explicitly citing the Gameplay bricks model. Regarding Ph.D. theses, we have identified 16 in English, 2 in Spanish, 20 (including HDRs) in French, and 2 in Portuguese. We conducted a detailed analysis of the metadata of the corpus, which we present below in the various tables.

Table 3. Distribution of the Number of Articles by Year.

Years	Number of Articles	Percentage
2008	1	2.1 %
2009	3	6.4 %
2010	3	6.4 %
2011	6	12.8 %
2012	4	8.5 %
2013	3	6.4 %
2014	3	6.4 %
2015	9	19.1 %
2016	11	23.4 %
2017	3	6.4 %
2018	1	2.1 %
Total	**47**	**100%**

Table 4. Distribution of the Number of Articles by Discipline.

Discipline	Number of Articles	Percentage
Computer Science	23	48.9%
Education	8	17.0%
Art	6	12.8%
Information Sciences	5	10.6%
Industrial Engineering	1	2.1%
Management	1	2.1%
Language / Literature	1	2.1%
Philosophy	1	2.1%
Health	1	2.1%
Total	**47**	**100%**

Table 5. Distribution of Authors by Nationality.

Nationality	Number of Authors
Germany	6
Belgium	3
Brazil	1
Bulgaria	5
Canada	4
Korea	2

Denmark	1
Spain	2
Estonia	1
France	19
Greece	1
Ireland	5
Italy	1
Japan	4
Netherlands	1
Portugal	1
Morocco	12
Mexico	5
United Kingdom	9
Russia	3
Sweden	3
Switzerland	1
Taiwan	4
USA	3

Table 3 illustrates the temporal coverage of our corpus, spanning the period from 2008 to 2018. Table 4 provides information about the disciplines identified based on the journals or conferences in which the scientific articles were published. Another aspect we considered relevant was the identification of authors and their nationalities. The data obtained is presented in Table 5.

The Gameplay Bricks Full Text Dataset

The corpus is primarily composed of PDF documents. These were converted into text for an initial preprocessing phase, enabling the analysis of the language used in scientific articles, PhD theses, and habilitation theses that were identified during our bibliographic research. The corpus used in this study consists exclusively of scientific articles, based on the full-text content extracted from PDF documents. The analysis of PhD theses will be addressed in future research and will be discussed in the context of the creation of new knowledge.

GROBID is a machine learning library designed to extract, parse, and restructure raw documents, such as PDFs, into structured XML/TEI documents. It is particularly suited for processing technical and scientific publications. In our study, we utilized the GROBID Web API, which provides a straightforward and efficient interface to the tool. The service was deployed within a Docker container running on Linux. For processing documents, we used the associated Python client, enabling concurrent processing of a batch of PDF files located in a specified directory. The experiments were conducted on a machine featuring an Intel® Core™ i7-4790K (8 threads) processor and 32 GB of RAM. No specific optimizations were applied to the GROBID processing pipeline, as the corpus size did not warrant such measures. GROBID was configured to generate TEI files with options tailored to the needs of our study. Specifically, the tool was set to perform sentence segmentation in the TEI XML output. This segmentation leverages the OpenNLP sentence detector, which is

recommended for scientific articles. The TEI generated by GROBID establishes a link between citation contexts and bibliographic references, enabling the construction of a matrix of relationships between citing and cited references. This approach allows us to connect the semantic categories produced by humans and machines in a network. The network will be a directed graph, with a label corresponding to the semantic category. For that purpose, we used Gephi to propose a practical case of visualization of the semantic network of games play from the labelled corpus Bastian, Heymann, and Jacomy (2009).

The following Table 6 shows a summary of the data processing produced by GROBID and corrected to produce a multilingual dataset of citation contexts to be explored. Indeed, the multilingual aspect poses difficulties in the conversion to TEI. We had to make corrections to improve context coverage. Nevertheless, the corrections we have made enable us to build a dataset referencing the founding articles of games bricks, and consequently to propose the dataset desired by our approach.

Table 6. Distribution of Articles, Citation Contexts, and References by Language .

Languages	Number of Articles	Number of Processed Articles	Number of Citation Contexts	Number of References in the Reference Corpus
English	44	44	962	58
Korean	2	2	49	3
Spanish	1	1	30	1
French	20	20	540	28
Indonesian	2	2	78	n.d.
Persian	1	n.d.	n.d.	n.d.
Portuguese	1	1	26	2
Russian	3	3	165	4
Swedish	1	1	68	1
Thai	1	n.d.	n.d.	n.d.
Total	**76**	**74**	**1918**	**97**

◈ *n.d.* not determined

Gameplay Bricks Labeling: A Human-Centric Perspective

The human approach was conducted in April 2018, relying in particular on Google Scholar via the use of the keywords “Brique Gameplay” and “Gameplay Brick” by associating the names “Djaouti” or “Alvarez” (Alvarez, 2018: pp42-73). A recent search carried out in 2023, again using Google Scholar, revealed 33 additional references for the same period. The corpus studied with the human approach thus represents 79 documents. The documents are then classified according to the type of

citation: The expert defined three labels in order to respond to his problem without taking into account existing categories: "Neutral", "Critic" and "Appropriation" (cf. Table). The Critic and Appropriation criteria are not mutually exclusive. Indeed, appropriation does not necessarily mean that the author expresses no criticism of the model. Some authors, like Pierre-Yves Hurel, take the trouble to criticize the model in order to appropriate it later on: To establish our own typologies (types of actions, types of rules), we propose to present and criticize the theory of gameplay bricks. As we shall see, this concept, which was created with the aim of improving game classification, can give us the tools we need for ideological analysis (Hurel, 2011, p29). With this in mind, it is worth drawing criticism also from the writings of researchers who have appropriated the model.

Cases of Gameplay Brick appropriation

In this subsection, the idea is to present the different types of appropriation of Gameplay Bricks identified by researchers and presented in Table 7. A dozen articles present an appropriation among which four types of appropriation can be identified. Five types of appropriation of the Gameplay Bricks model were identified in 2018. We'll take a closer look at these different types in the following subsections.

Table 7. Categorization of labels in the context of game bricks by a human expert, based on their purpose and knowledge of the field.

Labels	*Description of labels in the context of games bricks*	
Neutral	means that the Gameplay Bricks are merely cited by the article, but the author expresses no opinion on the model	
Critic	indicates that the article will significantly point out limitations or a disagreement with the model	
Appropriation	denotes a consideration of the model in the author's work.	
	Type 1 (T1)	**Use model:** Appropriation concerns the use of Gameplay Bricks to design or deconstruct Serious Games or video games.
	Type 2 (T2)	**Inspire methodologies:** Identified appropriation draws inspiration from the Gameplay Bricks to build new methodologies.
	Type 3 (T3)	**Integrate model:** The appropriation identified represents the integration of the Gameplay Bricks model into other models.
	Type 4 (T4)	**Develop experiments:** The appropriation of Gameplay Bricks is linked to the development of scientific experiments. However, in a more recent analysis of the additional references identified since 2023, we have identified a 5th type
	Type 5 (T5)	**Justifying a theoretical approach:** The appropriation of Gameplay Bricks is linked to a theoretical construction.

Appropriations of type 1: Use model

The first type of appropriation identified in the research literature concerns the use of Gameplay Bricks to deconstruct existing Serious Games or video games, or to help the design of new ones. This is the intended use when the model was developed. In this respect, we refer in particular to the article by Carlos Delgado-Mata, Ricardo Ruvalcaba-Manzano, Oscar Quezada-Patino, Daniel Gomez-Pimentel and Jesus Ibanez-Martinez: For the video game developed for this research, the bricks of interest are Move, Avoid and Reach. These types of bricks are well suited to our objective of developing a game that measures and develops fine and gross motor skills (Delgado-Mata, Ruvalcaba-Manzano, Quezada-Patino, Gomez-Pimentel, & Ibanez-Martinez, 2009, p5).

Type 2 appropriations: Inspiring methodologies

The second type of appropriation identified encompasses work that draws inspiration from the Gameplay Bricks to build methodologies. Marion Coville explains how she appropriated the Gameplay Bricks to build her experimental methodology for studying issues of gender, representation and role in video games (Coville, 2011, p 165). The researcher explains:

My methodology is based on this classification. First of all, I list the rules and actions available in the games, as well as the objectives and relationships to the world and universe in which the character evolves. I do this through my own experience of the game, while paying particular attention to the testimonies of other players. Once the modalities of interaction between the game and the player have been identified, I turn to the representation of heroines (Coville, 2011, p 172).

Type 3 appropriations: Integrating models

The third type of appropriation identified represents the integration of the Gameplay Bricks model into other models. This is the case, for example, of Yuri Gomes Cardenas, who proposes an ontology model designed to represent Serious Video Games. Among the elements that make up his model, the Gameplay Bricks model is thus mobilized (see Cardenas et al., 2014, p85)

Type 4 appropriations: Designing experiments

The fourth type of appropriation is linked to the development of scientific experiments. This is the case of Gaël Gilson, who proposed an experiment to study whether a gamer's virtual experience could represent an informal learning situation. One of the aims of the protocol was to ask subjects to identify the Gameplay Bricks they thought they would mobilize during the video-game activity, in order to understand how they ultimately they fit into the activity and the links they might establish with potential learning. The part of the protocol that calls upon the Gameplay Bricks is initially explained in the form of texts that are comprehensible to young subjects (Gilson, Draelants, Jardon, & Servais, 2016, p186). Once the subjects have been interviewed, the data collected is mapped (Gilson et al., 2016,

p187). Gameplay bricks are then listed in the same way as in the original English-language model, in the column Gameplay bricks employed.

Type 5 appropriations: Justifying a theoretical approach

This fifth type of appropriation aims to mobilize the Gameplay Bricks model to conduct a theoretical demonstration or corroborate theoretical approaches. This proposal does not intend to classify games, but to catalogue elements within a hierarchical structure. This catalogue can be used to describe the game according to its design space. It can also work as a framework to explore research questions related to games and gameplay, as proposed by the gamebricks classification, or to construct a vocabulary for describing, analyzing and critiquing games.

Results

Discipline and positioning overview

Based on the data listed in Table 2, Table 4 shows, in four columns, the total number of articles listed between 2009 and April 2018, the disciplines in which the Gameplay Bricks-related works were published, the total number of authors involved and their nationality, and finally their position with regard to the model. Table 4also presents the results in percentage terms. Overall, the model's diffusion is international, with France as the main country accounting for 20%. The main discipline to use the model is computer science (49%), followed by educational science (17%), art (12%), technology (12%) and CIS 10%, Critical feedback on the model accounts for the smallest percentage, 19%, behind 23.5% appropriations and a large majority of authors remaining neutral at 55.5%.

Distribution of critical positions and appropriations

Based on the data presented in Tables 2 and 3, Table 4 has been constructed to provide a more detailed breakdown of critical and appropriation stances regarding the Gameplay Bricks model. At this stage, the neutral stance has been excluded, as it does not enable the evaluation of the model. Table 3 reveals that authors from ten countries have adopted the Gameplay Bricks model, with over half of these countries being European. Conversely, authors from seven countries, more than half of which are also European, have expressed critical views of the model. From a disciplinary perspective, Communication and Information Sciences (CIS) emerges as the field with the highest level of appropriation, accounting for 30%. In contrast, Computer Science leads in terms of critical perspectives, with a rate of 50%. These findings now call for a closer examination of the nature of both appropriation and criticism, in order to rigorously evaluate the Gameplay Bricks model.

It is now time to see whether, on the one hand, other types of appropriation could be identified and, on the other, whether the set could give rise to an evaluative basis for situating its contribution to the Research.

Table 8 provides an overview of critical citations related to game brick models between 2008 and 2018. It highlights the multidisciplinary nature of research on this topic, spanning fields such as computer science, philosophy, design research and art.

This analysis reveals that computer science has the highest number of citations, reflecting its central role in the development and application of game brick models. These citations come from several countries, including Spain, Japan, the UK and the USA. We also note that most of the contributions in this field are in English, underlining the predominance of English as the main language for disseminating research on game brick models. This study demonstrates the broader theoretical and creative implications of game brick patterns with citations from philosophy, design research and art. Philosophy-related citations come notably from the USA, while design research is represented by an Australian study and art-related studies come from Canada and the Netherlands. Interestingly, while most of these publications are in English, one citation in the art category is in French, highlighting a certain linguistic diversity in the field.

In 2018, the different types of criticism identified are divided into 8 types and seem to be specific to each author: (2018, pp61-73):

Type 1: Misuse of Propp; Type 2: Subjective approach; Type 3: Lack of formalism; Type 4: Impossible classification; Type 5: Missing Meaning Bricks; Type 6: Means bricks irrelevant; Type 7: Distinguishing obligations and prohibitions; Type 8: Structure of games not studied.

In 2024, with the reading of the additional elements of the corpus, we can add a 9th type which would correspond to a formalism preventing the taking into account of storytelling or aesthetic.

Table 8. Critical Citation of Game Bricks models in the scientific literature from 2008 to 2018.

Critical Citation for Game Brick Models						
Discipline	T1-T8	Nationality	Language	Year	Nb. Aut	References
Computer Science	T3	Spain	English	2009	2	Reyno, E. M., & Cubel, J. A. (2009)
	T8	Japan	English	2010	4	Kim, T., [...] & Kondo, K. (2010)
	T9	United Kingdom	English	2015	2	Heintz, S., & Law, E. L. C. (2015)
	T4	USA	English	2015	5	Parkkila, J. [. . .] & Radulovic, F. (2015)
Philosophy	T4+T9	USA	English	2012	1	Thomas, L. D. (2012)
SIC	T2	USA	English	2008	1	Pennell, B. B. (2008)
	T5+T7	France	French	2011	1	Hurel, P. Y. (2011)
Design Research	T4	Australia	English	2017	2	Goddard W. & Muscat, A. (2017)
Art	T1+T4	Canada	French	2011	1	Arsenault, D. (2011)
	T6	Netherlands	English	2011	1	Veugen, J. I. L. (2011)
		Canada	English	2017	1	Therrien, C. (2017)

Table 9 provides an overview of critical citations of Game Brick models across various academic disciplines between 2008 and 2018. These citations indicate an analytical or evaluative engagement with Game Brick models rather than neutral references. The majority of critical citations appear in Computer Science, with six publications from diverse national backgrounds, including France, Morocco, the USA, the Netherlands, and Germany. The linguistic diversity of these citations is also notable, with publications in English, French, and German, reflecting the global discourse surrounding Game Brick models. Beyond Computer Science, critical assessments of these models are present in Science of Information and Communication (SIC) (Taiwan, 2013), History (Germany, 2011), Management (France, 2012), Education (South Korea, 2013), and Economy (Sweden, 2009). These publications are written in English, French, Korean, and Swedish, underscoring the multilingual engagement with Game Brick models in academic research. The number of authors per publication varies, from single-authored works to multi-author collaborations, suggesting different approaches to critical analysis across disciplines. The temporal distribution of these citations highlights key years of critical engagement, particularly in 2009, 2013, and 2015, indicating sustained but irregular scrutiny of the models. The presence of critical citations across multiple fields demonstrates the interdisciplinary impact of Game Brick models, with researchers actively assessing their theoretical, methodological, and practical implications.

Table 9. Neutral Citation of Game Bricks models in the scientific literature from 2008 to 2018.

Critical Citation for Game Brick Models					
Discipline	Nationality	Language	Year	Nb. Aut	References
Computer Science	France	English	2009	3	Carron, T. [...] & Mangeot, M. (2009)
	France	French	2010	1	Muratet, M. (2010)
	Morocco	English	2014	2	El Borji, Y., & Khaldi, M. (2014)
	USA	English	2015	2	Schatz, K., & Riippel, U. (2015)
	Netherland	English	2016	1	Carvalho, B., M. (2016)
	Germany	Deutch	2017	1	Piepr, J. (2017)
SIC	Taiwan	English	2013	4	Yang, H. T., [...] & Chen, K. T. (2013)
History	Germany	English	2011	1	Goelz, C. (2011)
Management	France	French	2012	3	Chollet, A., [...] & Rodhain, F. (2012)
Education	South Korea	Korean	2013	2	Kwon, C. S., Woo, T. (2013)
Economy	Sweden	Swedish	2009	1	Ahmet, Z. (2009)

Table 10 presents an overview of appropriation citations of Game Brick models across multiple academic disciplines from 2008 to 2018, classified into five subcategories (T1–T5). Appropriation citations indicate instances where researchers have integrated, adapted, or extended the Game Brick models within their work rather than merely analyzing or critiquing them. The dataset spans a wide range of fields, including Computer Science, Philosophy, Health, Management, Science of Information and Communication (SIC), Education, Language Sciences, Architecture, Design Research, and Art. The Computer Science domain exhibits the highest number of appropriation citations, with contributions from Estonia, the United Kingdom, Brazil, Italy, Germany, and Sweden, predominantly in English and Portuguese. The temporal distribution highlights increased adoption in 2014, 2015, 2017, and 2018, with author teams ranging from single to multi-author collaborations (up to eight contributors per study). This suggests a progressive incorporation of Game Brick models into computational frameworks and technological innovations. Beyond Computer Science, Philosophy (Portugal, 2016) and Health (France, 2012 and 2016) display instances of appropriation, primarily in English and French, focusing on conceptual and applied methodologies. Management (Germany, 2011) also features an English-language appropriation citation, reflecting its relevance in organizational and strategic domains. The field of Science of Information and Communication (SIC) includes citations from Mexico, Belgium, and Denmark (2009–2015), highlighting a multilingual engagement (English and French) and a growing interest in the theoretical adaptation of Game Brick models. Similarly, Education (Germany, Belgium, Russia, 2015–2016) demonstrates a diverse linguistic profile (English, French, and Russian), emphasizing the use of Game Brick models in pedagogical and instructional design. Other disciplines, including Language Sciences (France, 2016), Architecture (Turkey, 2013), and Design Research (Singapore, 2013), show targeted appropriation, indicating the versatility of these models across different research fields. Finally, Art (France, Sweden, 2011–2014) exhibits an engagement with both theoretical and applied perspectives, reinforcing the interdisciplinary impact of Game Brick models. The temporal distribution of appropriation citations reveals a steady adoption pattern, with peaks in 2014, 2015, and 2016, reflecting a maturing research interest in integrating Game Brick models into diverse disciplinary frameworks. The presence of multilingual publications and global contributions underscores the broad academic reception and adaptability of Game Brick models, reinforcing their significance as a foundational tool in various research fields.

Table 10. Appropriation Citation of Game Bricks models in the scientific literature from 2008 to 2018 with subdivision T1 to T5.

Appropriation Citation for Game Brick Models						
Discipline	T1 - T5	Nationality	Language	Year	Nb. Aut	References
Computer Science	T1 T3 T3 T1 T3 T1 T5 T1	Estonia United Kingdom Brazil Brazil Italy Germany Brazil Sweden	English English Portugese Portugese English English English English	2010 20122014 2014 2015 2017 2018 2018	1 3 1 4 8 3 3 1	Henno, J. (2010) Carter, C., [...] & Hartley, T. (2012) Murakami, L. C. […] & Almeida Macedo, D. (2014) Carvalho, B., M. , [...] & Rauterberg, M. (2015) Schmidt, S., [...] & Möller, S. (2017) Dominguez, R.G. [...] & Oliviera Venâncio, R.D. (2018) Laine, T. H. (2018)
Philosophy	T2 T2	Portugal Portugal	English English	2016 2016	1 2	Cardoso, P. J. C. (2016) Cardoso, P. & Carvalhais, M. (2016)
Health	T1 T1	France France	French French	2012 2016	3 1	Mader, S. […] & Levieux, G. (2012) Ben-Sandoun, G. (2016)
Management	T1	Germany	English	2011	4	Duin, H. […] & Thoben, K-D. (2011)
SIC	T1 T3	Mexico Belgium	English French	2009 2011	5 1	Delgado-Mata, C., [...] & Ibanez-

	T2 T2	Belgium Denmark	French English	2012 2015	1 1	Martinez, J. (2009) Hurel, P. Y. (2011) Palmieri, J. (2012) Otzen, T. (2015)
Education	T1 T4 T3	Germany Belgium Russia	English French Russian	2015 2016 2016	3 1 3	Müller, B. C., Reise, C., & Seliger, G. (2015) Gilson, G. (2016) Akchelov E.O.[...] & Nikitina K.S. (2016)
Language Sciences	T5	France	French	2016	1	Schmoll, L. (2016)
Architecture	T1	Turkey	English	2013	1	Örnek, M.A. (2013)
Design Research	T1	Singapore	English	2013	2	Yen C.C. & Lee J.M. (2013)
Art	T2 T1 T4	France France Sweden	French French English	2011 2014 2014	1 1 1	Coville, M. (2011) Fernandez, M.M. (2014) Ghys, K. (2014)

Disciplines, Citation Types, Languages, and Countries

We provide an overview of citation contexts that are not in English, highlighting their linguistic diversity and their relevance to the research. The annotation process for sentences containing citation contexts is detailed in the Table 6, where these contexts are categorized by language and corresponding annotations. Additionally, we include the translations employed during the labeling process to ensure consistency and accuracy across languages. This approach allows us to illustrate the multilingual nature of citation contexts while maintaining a standardized framework for analysis and interpretation.

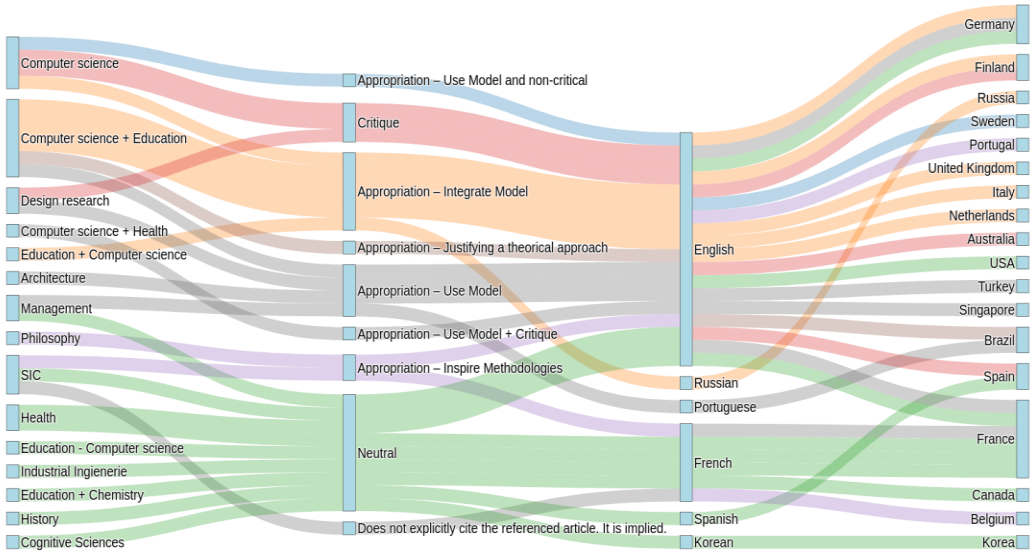


Figure 1. Citation Flow of Game Bricks Models Across Disciplines, Citation Types, Languages, and Countries.

Categories generated by the prompting approach

Figure 2 illustrates the distribution of citation context typologies within the Game Bricks research corpus during the 2008–2018 period. The data highlights the predominance of certain typologies, indicating recurring conceptual frameworks in the field, as suggested by the "Definition" and "Appropriation" categories. The dominant paradigm is thus definition and appropriation. Neutrality and criticism are more difficult to capture with our approach based on the produced sample.

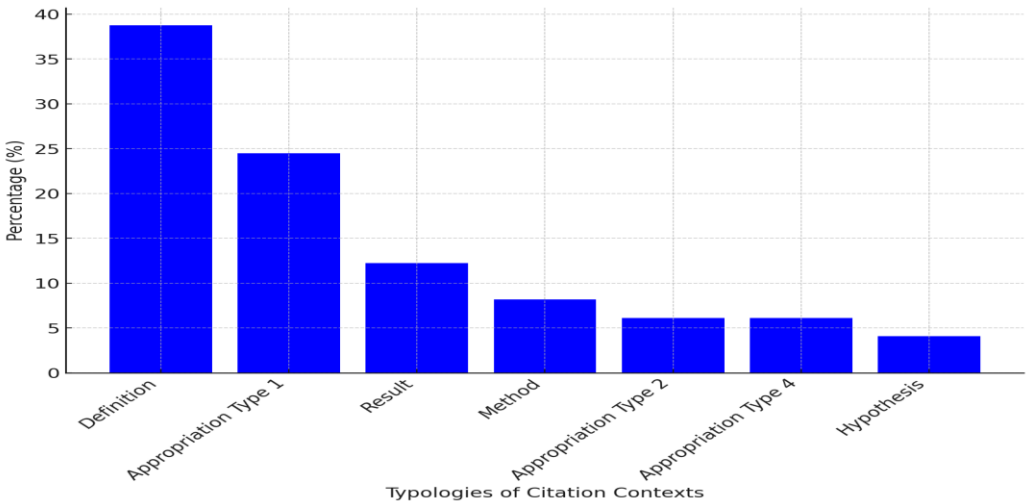


Figure 2. Distribution of Citation Context Typologies in Game Bricks Model (2008-2018).

Discussion of Experimental Results

Limitations of Our Study

The first challenge lies in the consolidation of the dataset, with a major constraint related to its multilingual dimension. Current tools do not yet offer a multilingual approach for corpus processing. As a result, its consolidation relies on human-based and time-consuming methods. For instance, Persian, Thai, and even Korean corpora could not undergo the segmentation stage, which is crucial for generating the attributes that link citation contexts to bibliographic references. The second limitation concerns the processing and assignment of attributes to citations within the text. We observed that the attributes used to associate citation segments with references are often incorrect. While this does not prevent annotation—since the citation context segment is extracted—it does hinder the ability to accurately link it to references. Moreover, handling multiple references remains challenging for this type of processing. The third limitation is the lack of adherence to citation standards in some papers, leading to processing errors. The fourth difficulty concerns the design of prompts and the reproducibility of results.

Perspectives

Despite these challenges, the approach using LLMs and prompts remains promising, provided that we can generate prompts based on categories that eliminate any semantic or conceptual indeterminacy. This is likely the next step in improving results with this type of approach. During this study, it was interesting to allow the system to propose multiple annotations for a given citation context. Granting this flexibility enabled broader coverage and improved system-generated annotations. We will focus on new reasoning models, with a particular emphasis on Chain of Thought approaches, which yielded promising results in our experiments. Indeed, the *Chain of Thought* approach will enable the explicit structuring of reasoning by breaking down a task into several intermediate steps. In citation analysis, this will allow for a better distinction between the different functions of a citation, especially in cases where citation contexts may be ambiguous. Finally, stabilizing our input corpus will allow us to conduct an evaluation comparing AI-based annotation with human annotators using the Kappa coefficient. Finally, the stabilization of our input corpus will enable us to perform an evaluation comparing AI-based annotation with human annotators using the Kappa coefficient. To this end, we will compare several llm's using tools such as LMStudio.

Conclusion

This study presents an in-depth analysis of citation contexts surrounding the gameplay bricks model between 2008 and 2018, comparing human expert analysis with AI-assisted approaches. Our results highlight both the potential and limitations of AI-assisted citation context analysis, thus emphasizing the need for hybrid approaches that integrate human expertise with machine learning capabilities.

One of the main findings of this study is the predominance of definition and appropriation categories across different disciplines, illustrating the widespread

adoption of the Gameplay Bricks model. The data reveals that computer science fields tend to appropriate this model for practical applications, while humanities and social sciences engage with it more critically. These variations highlight the influence of disciplinary conventions on citation practices and suggest that citation contexts are shaped by epistemic cultures that determine how knowledge is referenced, criticized, and integrated. Our analysis reveals the international and multidisciplinary impact of the gameplay bricks model, with citations spanning nine languages and multiple academic fields. Computer science emerges as the primary field of application (48.9%), followed by education (17.0%) and arts (12.8%), thus demonstrating the model's broad relevance. Temporal analysis shows adoption peaks in 2015-2016, suggesting a maturation phase in the model's development and application.

A second aspect concerns our methodological approach, which combines human annotation and AI-assisted classification through prompt engineering, highlighting the potential for large-scale automated citation analysis. ChatGPT-generated analyses offer advantages in terms of scalability and efficiency, enabling the processing of extended multilingual corpora that would be very time-consuming for human annotators. However, AI's ability to capture nuanced critiques and neutral citations remains limited. This limitation becomes even more pronounced when considering the expert-driven categorization of inherent critiques of Game Brick models. A detailed analysis reveals eight distinct types of criticism. Type 1 critiques highlight the erroneous application of Propp's framework (2018, p. 58), where studies misinterpret or misapply narrative structures. Type 2 critiques address a subjective approach (2018, p. 60), pointing out a lack of methodological rigor and an overreliance on interpretation. Type 3 critiques emphasize a lack of formalism (2018, p. 62), indicating that some applications fail to adopt a structured theoretical framework. Type 4 critiques argue that the model leads to an impossible classification (2018, p. 63), suggesting that its structure does not allow for a coherent categorization of game elements.

Further critiques focus on the content of Game Brick models. Type 5 critiques identify missing "Means Bricks" (2018, p. 65), arguing that essential intermediary elements necessary for game mechanics are absent. Conversely, Type 6 critiques question the relevance of certain "Means Bricks" (2018, p. 65), indicating that some components do not meaningfully contribute to game design. Type 7 critiques stress the need to differentiate obligations from prohibitions (2018, p. 66), underscoring a conceptual gap in distinguishing required actions from restricted ones. Finally, Type 8 critiques highlight the lack of analysis of game structures (2018, p. 67), pointing to a broader limitation in addressing overarching game frameworks. These identified critique categories offer a more nuanced and structured understanding of the scientific discourse surrounding Game Brick models. They emphasize not only theoretical and methodological gaps but also practical issues in the application of the framework, underscoring the need for further refinement and conceptual clarity.

This limitation results from both model biases and the inherent complexity of interpreting citation contexts, which often require deep domain expertise and understanding of implicit rhetorical subtleties. This finding aligns with previous

research on LLM capabilities in academic discourse analysis. A notable limitation of our study lies in the multilingual nature of the dataset. Current AI tools, including ChatGPT, still struggle with complex linguistic variations, particularly for underrepresented languages. While citation contexts in English and French were processed with relatively high accuracy, languages such as Persian, Thai, and Korean posed challenges due to insufficient training data and segmentation difficulties. Future research should focus on refining multilingual NLP models to better capture citation contexts across various linguistic environments. Furthermore, the gaps between human and AI-generated annotations highlight the need for more robust prompting strategies. Our results indicate that few-shot learning and chain-of-thought approaches improve AI citation classification accuracy but still cannot fully replicate human interpretative capabilities. The observed inconsistencies suggest that prompt refinement is essential for optimizing AI performance in citation analysis. The methodological challenges encountered, particularly in multilingual processing and prompt engineering, highlight important areas for future research, including:

- Developing more robust tools for multilingual citation context processing
- Improving reference linking accuracy in complex citation networks
- Refining prompt engineering techniques for specialized academic discourse
- Creating standardized evaluation frameworks for citation context analysis

In conclusion, this research contributes to the debate on AI-assisted citation analysis by proposing a comparative study spanning multiple languages and disciplines. Based on a case study, we have produced a corpus of citation contexts related to the Gameplay Bricks framework, along with prompts to categorize these contexts. We also provide a dataset of contexts annotated by an expert. Additionally, we propose a methodology for implementing categorization through prompts. It illuminates both the opportunities and challenges associated with using AI to interpret citation contexts, advocating for more sophisticated tools capable of accounting for linguistic and disciplinary variations. Moving forward, the development of improved multilingual NLP models and refinement of AI citation categorization techniques will be essential for enhancing the reliability and applicability of citation context analysis in academic research.

Acknowledgments

This work was supported by ANR-20-CE38-0003-01. The authors declare that they have no conflict of interest. This manuscript has been prepared with the assistance of artificial intelligence (AI)-based tools to support the writing process. Specifically, AI was employed for language refinement, grammar correction, and structuring of certain sections. However, all intellectual contributions, critical analysis, and interpretations presented in this work remain the sole responsibility of the authors.

References

Aithal, P., & Aithal, S. (2023). Application of chatgpt in higher education and research–a futuristic analysis. International Journal of Applied Engineering and Management Letters (IJAEML), 7 (3), 168–194.

Alvarez, J., Djaouti, D., Ghassempouri, R., Jessel, J.-P., Methel, G. (2006). Morphological study of the video games. Actes du colloque "cgie 2006". Perth, Australie.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session* (pp. 81-87).

Bastian, M., Heymann, S., Jacomy, M. (2009). Gephi: An open source software for exploring and manipulating networks.

Bertin, M., & Atanassova, I. (2024). Linguistic perspectives in deciphering citation function classification. *Scientometrics*, 1-13.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? a review of studies on citing behavior. Journal of documentation, 64 (1), 45–80.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., . . . Amodei, D. (2020a). Language models are few-shot learners. H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), Advances in neural information processing systems (Vol. 33, pp. 1877–1901).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Cabanac, G., Labbé, C., Magazinov, A. (2021). Tortured phrases: A dubious writing style emerging in science. evidence of critical issues affecting established journals. arXiv preprint

Cabanac, G., Labbé, C., Magazinov, A. (2022). Bosom peril" is not "breast cancer": How weird computer-generated phrases help researchers find scientific publishing fraud. Bulletin of the Atomic Scientists, January, 13.

Cabanac, G., Labbé, C., Magazinov, A. (2022). The 'problematic paper screener' automatically selects suspect publications for post-publication (re)assessment. Retrieved from https://arxiv.org/abs/2210.04895

Cardenas, Y.G., et al. (2014). Modelo de ontologia para representação de jogos digitais de disseminação do conhecimento (Unpublished doctoral dissertation). Orientador: João Bosco da Mota Alves. Coorientador: Denilson Sell. Dissertação (mestrado)-Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, Florianópolis.

Chen, Q., Sun, H., Liu, H., Jiang, Y., Ran, T., Jin, X., . . . Chen, H. (2023). A comprehensive benchmark study on biomedical text generation and mining with chatgpt. bioRxiv .

Coville, M. (2011). Hello, i'm a woman gamer. i got tired of people telling me to get a life. i began studying video games instead. Bruxelles, Belgique.

Delgado-Mata, C., Ruvalcaba-Manzano, R., Quezada-Patino, O., Gomez-Pimentel, D., Ibanez-Martinez, J. (2009). Low cost video game technology to measure and improve motor skills in children. Africon, 2009. africon'09 (pp. 1–6). IEEE.

Dong, C., & Schäfer, U. (2011). Ensemble-style self-training on citation classification. In *Proceedings of 5th international joint conference on natural language processing* (pp. 623-631).

Else, H. (2021). Tortured phrases' give away fabricated. Nature, 596, 328–9. Gilson, G., Draelants, H., Jardon, D., Servais, O. (2016). L'expérience virtuelle des joueurs comme situation d'apprentissage informel. Université de Mons, Belgique.
Hernández-Alvarez, M., & Gomez, J.M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering, 22 (3), 327–349.
Hurel, P.Y. (2011). Analyse idéologique des jeux vidéo/une méthode ludonarrative pour les jeux mis en récits (Unpublished doctoral dissertation). Université de Liège, Liège, Belgique.
Jha, R., Jbara, A.-A., Qazvinian, V., Radev, D.R. (2017). Nlp-driven citation analysis for scientometrics. Natural Language Engineering, 23 (1), 93130.
Karan, B., & Angadi, G. (2023). Potential risks of artificial intelligence integration into school education: A systematic review. Bulletin of Science, Technology & Society, 43 (3-4), 67–85.
Khan, N.D., Khan, J.A., Li, J., Ullah, T., Zhao, Q. (2025). Leveraging large language model chatgpt for enhanced understanding of end-user emotions in social media feedbacks. Expert Systems with Applications, 261, 125524. https://doi.org/10.1016/j.eswa.2024.125524
Lahiri, A., Sanyal, D.K., Mukherjee, I. (2023). Citeprompt: using prompts to identify citation intent in scientific papers. 2023 acm/ieee joint conference on digital libraries (jcdl) (pp. 51–55).
Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1-35.
Lyu, D., Ruan, X., Xie, J., Cheng, Y. (2021, April). The classification of citing motivations: a meta-synthesis. Scientometrics, 126 (4), 3243–3264.
Martel, E., Lentschat, M., Labbé, C. (2024). Detection of tortured phrases in scientific literature. arXiv preprint arXiv:2402.03370
Mathebula, M., Modupe, A., Marivate, V. (2024). Chatgpt as a text annotation tool to evaluate sentiment analysis on south african financial institutions. IEEE Access.
Nishikawa, K., & Koshiba, H. (2024). Exploring the applicability of large language models to citation context analysis. *Scientometrics*, *129*(11), 6751-6777.
Radford, A. (2018). Improving language understanding by generative pretraining.
Tahamtan, I., & Bornmann, L. (2022). The Social Systems Citation Theory (SSCT): A proposal to use the social systems theory for conceptualizing publications and their citations links. *Profesional de la información*, *31*(4).
Teufel S, Carletta J, Moens M (1999) An annotation scheme for discourse-level argumentation in research articles. In: Henry S. Thompson AL (eds) Proceedings of the ninth conference on European chapter of the association for computational linguistics, Bergen, Norway, 08–12 June 1999. Association for Computational Linguistics, pp 110–117.
Teufel S, Siddharthan A, Tidhar D (2006) Automatic classification of citation function. In: Mirella Lapata HTN (editor) Proceedings of the 2006 conference on empirical methods in natural language processing, Sydney, Australia, 25–27 October 2006. Association for Computational Linguistics, 1610091, pp 103–110
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, *35*, 24824-24837.

Zhang, B., Fu, X., Ding, D., Huang, H., Dai, G., Yin, N., ... & Jing, L. (2023). Investigating chain-of-thought with chatgpt for stance detection on social media. *arXiv preprint arXiv:2304.03087*.

Zhang, Y., Wang, Y., Wang, K., Sheng, Q. Z., Yao, L., Mahmood, A., ... & Zhao, R. (2023). When Large Language Models Meet Citation: A Survey. *arXiv preprint arXiv:2309.09727*.

https://doi.org/10.51408/issi2025_066

Co-funding Networks as a New Tool in Research Evaluation: A Linked Open Data-Based Study of the Seventh Framework Programme Projects

Niliek Silva-Alés[1], Antonio Perianes-Rodríguez[2]

[1]*nisilvaa@bib.uc3m.es*, [2]*antonio.perianes@uc3m.es*
Universidad Carlos III de Madrid, ROR, Department of Library and Information Science. DIVALab Group, C/ Madrid, 128, 28903 Getafe (Madrid) (Spain)

Abstract

There is a growing interest in studying the influence of funding on scientific progress. Through exploration of the connections between funding acknowledgements (FAs), which link research results to funding sources, science communication processes can be understood and their influence in the international context can be evaluated. Such analyses become more complex when the projects involved have two or more funding sources. This study examines FAs that mention the Seventh Framework Programme (FP7) and tries to achieve a broader, fuller, more singular view than previous studies of FP7 by visualising co-funding networks and conducting a structural analysis of inter-agency relationships. This is done using open sources that have been linked after exhaustive data cleansing and harmonisation and the assignment of unique identifiers. Compliance with the objectives of the three most visible, most productive programmes is also examined, and the geographical distribution of the agencies participating in co-funding networks is evaluated. One intriguing result shows that the number of projects with associated publications has risen 21% thanks to FAs. Considerable differences between programmes are also revealed: IDEAS-ERC is the programme with the highest number of co-funders, and PEOPLE is the programme with the densest, most cohesive network. Lastly, it is found that a stronger commitment is required from all the actors involved in the course of co-funding and publication to ensure that the funding data provided is of the right quality to facilitate accurate, transparent, useful, full evaluations.

Introduction

Funding acknowledgements (FAs) generally occupy a section of their own in scientific articles, listing all the people and organisations that have funded, supported or contributed to the paper (Wang & Shapira, 2011). FA information is essential for understanding the research context, its communication processes and the essential role played by funding in scientific advancement. Information of this sort lends itself to various types of analysis, including the creation of co-funding maps, as a subset of scientific collaboration networks, with distinctive information that is useful for tracing other kinds of intellectual influences (Costas & van Leeuwen, 2012).

The quality of co-funding analysis is affected by the availability, integrity and quality of the metadata used and by the workableness of linking funding with published results for an accurate evaluation of the most efficient, effective funding systems, programmes and policies. FAs are crucial to such studies, because they name funding agencies and identify projects, and these are the basic components for building networks and establishing links between agents to connect funding with scholarly output. As the section on methodology will explain, the funding metadata used in this study were obtained from open sources that were combined to expand upon the

quantitative analysis perspective by adding the structural facet furnished by network analysis.

Furthermore, science policies in the European Union (EU) call for the transcension of traditional barriers to research. In that effort, they support transnational, multisector and multi, inter- and transdisciplinary research. Funding opportunities themselves, whether individual or collective, promote diverse, heterogeneous funding and funding co-use (Aagaard et al., 2021).

The Framework Programmes are a good example. The Framework Programmes are the main funding instrument for consolidating the European Research Area. The seventh programme (FP7) in particular made project co-funding one of its basic principles (European Commission, 2007). Funding plans for 2007-2013 were divided into collaborative projects, networks of excellence and coordination and support actions, with the objective of enhancing the competitiveness and excellence of science in Europe.

The main objectives of the Seventh Framework Programme are not limited to producing co-funding, but also include the following: to promote excellence in research, to foster competitiveness and economic growth, to help address social challenges, to strengthen human potential, to foster researcher mobility and to promote transnational cooperation in research. FP7's budget was 66% higher than FP6's. Eighty-one percent of the budget (44,600 million euros) was assigned to four preferred programmes, FP7-COOPERATION, FP7-IDEAS, FP7-PEOPLE and FP7-CAPACITIES (European Commission, 2018).

The main benefits of European funding as opposed to national funding are the following: access to international research, networking with leading scientists, better reputation, greater possibilities of obtaining additional funding and the formation of international consortia. The end result of all these efforts was greater participation by actors and stakeholders, helping to cast a more solid foundation for cooperation, at the national level as well.

FP7 is one of the few research funding programmes that maintained its budget, thus placing it in a better light in the eyes of the international research community. Global economic development no longer depends on the "triad" of North America, Japan and Europe. New actors are arising, including China, Korea and Latin-American countries, generating multipolar competition and creating the need to establish fresh partnerships (European Commission, 2018).

Lastly, prior studies of co-funding networks (Boyack, 2009; Wang & Shapira, 2011; Grassano et al., 2017; Aagaard et al., 2021; Mugabushaka, 2022 and Perianes-Rodríguez et al., 2024a) agree that the general processes involved in conducting these kinds of analyses are data gathering, data cleansing and data harmonisation. These processes vary depending on the underlying funding, its influence on the research and the way the funding is recorded. However, few studies run detailed analyses of the resulting networks and visualisations.

This study, then, examines the Seventh Framework Programme's co-funding network, bringing fresh perspectives that complement those described in other papers on FP7 funding (Mugabushaka, 2020; Ardanuy et al., 2023; 2024). Co-funding can help redefine traditional scientific collaboration practices, widen the

scope covered by the scarce economic resources available and underwrite projects that can make disruptive breakthroughs. In addition, because co-funding links diverse open data sources together, it enriches and expands the scope of accountability, helping to make the evaluation of science, technology and innovation policies more open and easily reproduced and fostering more efficient, more inclusive, more transparent evaluation ecosystems.

Objectives

The main objective is to run an open-source structural analysis of the effects of co-funding on FP7-funded projects and its role in scientific development, based on a study of research results published in scientific journals. A thorough empirical study explores the usefulness of the funding information reported in publication acknowledgements and the influence of co-funding, focusing especially on analysis of the resulting co-funding networks. For these purposes, the following secondary objectives are defined:

- To find the proportion of projects correctly labelled with their identifier in the FAs of papers published in scientific journals.
- To extract open-source funding metadata to determine their quality and the synergies that could result if they are appropriately combined.
- To determine the geographic composition of co-funding networks and to identify the main participating agencies.
- To analyse the relational indicators of the European funding programmes that have the highest number of projects with reported publications, to determine compliance with the funding programmes' main objectives.
- To identify the problems with co-funding data and the action needed to improve the quality of results based on metadata of this sort.
- To measure compliance with FP7's strategies and objectives on the basis of structural analysis of the published results of funded projects.

This study is structured as follows: "Data and methodology" describes how data were downloaded and processed and what methodology was used to create the co-funding maps; "Results" presents the base map of FP7 project co-funding and the leading bibliometric and structural data of the four target networks; "Discussion and conclusions" sums up the main findings on performance differences between the analysed programmes; lastly, "Limitations and future work" contains recommendations for improving the quality of data for co-funding analysis, proposes practical steps for the various agents involved and maps out future lines of research aimed at ascertaining the visibility and influence of co-funded papers.

Data and Methodology

The ties between research funding and the scientific results of funded research are hard to track and often require access to separate reports from researchers or funders (Wang & Shapira, 2011). Although FP7 project funding ended in 2014, the last funded projects were not complete until 2019, and papers reporting work funded by

FP7 projects are still being published today (Ardanuy et al., 2023). These are the results of research that needs to be analysed from a holistic perspective, making use of open data to gain a clearer picture of the synergies between funders and to determine the influence of the publications that funders sponsor.

From the start FP7 was split into four programmes, Cooperation, Ideas, People and Capacities, as a means of better achieving its European research support objectives. The main anticipated results included stronger industrial competitiveness for Europe, job growth and the identification of new ways to improve research and innovation infrastructure to ensure the quality of science and effective complementarity among Community institutions (European Commission, 2016).

Analysis of funding programmes based on the data available from open bibliographic sources can be used to evaluate the operation, scope and impact of these programmes and determine their efficacy and transformative ability. One source used in this study is CORDIS[1], which is the source of official FP7 data on projects and publications reported by beneficiaries (Ardanuy, 2023). Another data source is Crossref[2], the leading international registration agency of Digital Object Identifiers (DOIs), through its Open Funder Registry (OFR) initiative. Crossref is a complementary source that provides data on publication funding based on the information released by authors and editors in publication acknowledgements (Álvarez-Bornstein & Montesi, 2020). Authors and editors must furnish information on the funding agency involved, its unique identifier and the project's number (Kramer & de Jonge, 2022). To reach its objectives, this study uses the methodology described in Perianes-Rodríguez et al. (2024a), which employs linked open metadata from various data sources to analyse funding agencies' performance. Account is also taken of co-funding network studies described in Boyack (2009), Wang & Shapira (2011), Grassano et al. (2017) and Mugabushaka (2022), which are the theoretical and visual forefathers of this paper. The analytical processes are described below.

Data gathering and processing

Data cleansing and harmonisation require an immense amount of manual work to locate and enter information related with the target funding sources (Wang & Shapira, 2011). The first step in this project was to download data on projects and publications from CORDIS and data on the various FP7 programmes mentioned in acknowledgements in OFR, in July 2023. Next, the data were disambiguated and standardised. Of the 320,448 rows downloaded, 318,322 (99.33%) were disambiguated, and the funder's ISO 3166-1 alpha-3 country codes were added. In the case of funders with headquarters in more than one country, the country of the official headquarters was used. Of the 119,284 lines of European Commission funders, the 99,621 rows that included project numbers were reviewed. After harmonisation 91,887 rows (92.23%) were left. This intense cleansing process considerably boosted the quality and accuracy of the original data used in the structural analysis of the co-funding networks.

[1] https://cordis.europa.eu/

[2] https://www.crossref.org/services/funder-registry/

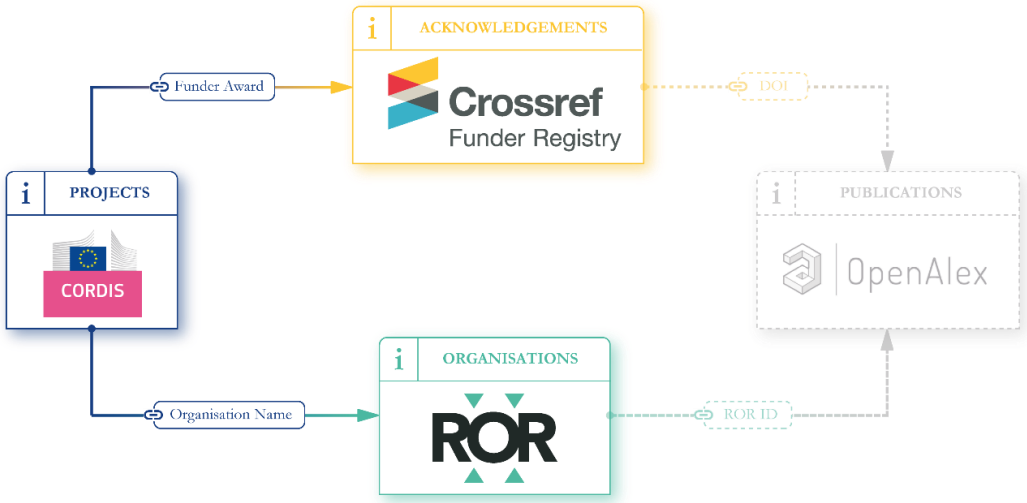


Figure 1. Linked open data. Schema of sources and normalised identifiers.

The Research Organization Registry (ROR) was also used to complete or correct institutional identifiers. This source provides standardised information about institutions and enables research organisations to be linked to their researchers and their research results (ROR, 2024). Figure 1 illustrates the linking procedure and the standardised identifiers used to connect the three data sources. OpenAlex is shown in grey, because it will be used in future research work.
The metadata extracted from each source are shown in Table 1.

Table 1. Sources and list of downloaded metadata.

Source	*Metadata*
CORDIS	Project identifier, title, publication DOI, total funding, year, grantee organisation, country, FP7 funding scheme.
OFR	Publication DOI, title, funder DOI, funder name, project identifier.

The main problems found in the information downloaded from OFR were disparities in funder names, gaps in the identification of project codes and an absence of essential data, like the funder's country. For example, the Dutch Research Council (NWO) appears under 150 variants of its name, and the Karolinska Institutet has 35 name variants listed.

Table 2. Basic indicators of projects and publications by source.

Sources	*Projects*
CORDIS	25,785
CORDIS with publications	14,297
CORDIS and OFR	9,250
CORDIS with publications and OFR	6,230
OFR only	3,020

Sources	*Publications*
CORDIS	216,004
OFR	47,493
CORDIS and OFR	7,333

Finally, 7,333 publications that matched in CORDIS and OFR could be connected. They referred to 9,250 projects, 6,230 of which had publications reported in CORDIS. Surprisingly, 3,020 projects were located without publications reported in CORDIS but with explicit acknowledgements in OFR, which is to say that one out of every five projects with publications was not included in CORDIS.

Structural indicators

The following structural indicators were analysed:

a) Nodes: Total number of funding agencies.
b) Edges: Number of connections between nodes.
c) Density: Proportion of real links relative to the maximum number of possible edges.
d) Average degree: The average number of edges per node.
e) Degree and betweenness centralisation: Centralisation of a network is a measure of how central its most central node is in relation to how central all the other nodes are. So, the measures analyse centralisation of degree (number of edges with adjacent nodes) and betweenness (frequency of a node on the shortest paths between other actors).
f) Average distance: Average shortest path length between nodes. It is a measure of the efficiency of communication in a network.
g) Diameter: The shortest distance between the two most distant nodes, that is, the longest of all the path lengths in the network.

Visualisation of co-funding networks

The networks were visualised using Pajek[3] (Batagelj & Mrvar, 2004). To create the networks, multiplicative counting (Perianes-Rodríguez & Ruiz-Castillo, 2015) and fractional counting (Perianes-Rodríguez et al., 2016) were employed. It was decided to use fractional counting because that is the method recommended in bibliometric studies of countries and research organisations (Waltman & van Eck, 2015). Analyses based on fractional counting show that scientific collaboration preferably takes place with national partners, and this circumstance helped in labelling the resulting clusters.

For the creation of the co-funder network base map, the methodology described by Leydesdorff & Rafols (2009) was used. Communities were extracted using the Louvain algorithm (Blondel et al., 2008). For spatial representation, the Kamada-Kawai algorithm (1989) was employed. Of the initial 4,459 funders, the analysis was restricted to the 947 that participated in the co-funding of at least 10 publications (not including EU funders). The national and regional ministries of each European country were grouped under a single government funder.

The aggregated data set is available as supplementary material at https://doi.org/10.5281/zenodo.14502483 (Perianes-Rodríguez et al., 2024b).

Results

The base map shows the general co-funding patterns of the set of projects and publications. Each of the nodes represents a funding agency. Node size depends on the number of papers co-funded. Links become thicker and darker as the number of co-funded publications increases. The base map contains 947 nodes linked by 29,521 edges (the IDEAS-ERC, PEOPLE and HEALTH co-funding maps are available in annexes 1 to 3).

[3] http://mrvar.fdv.uni-lj.si/pajek/

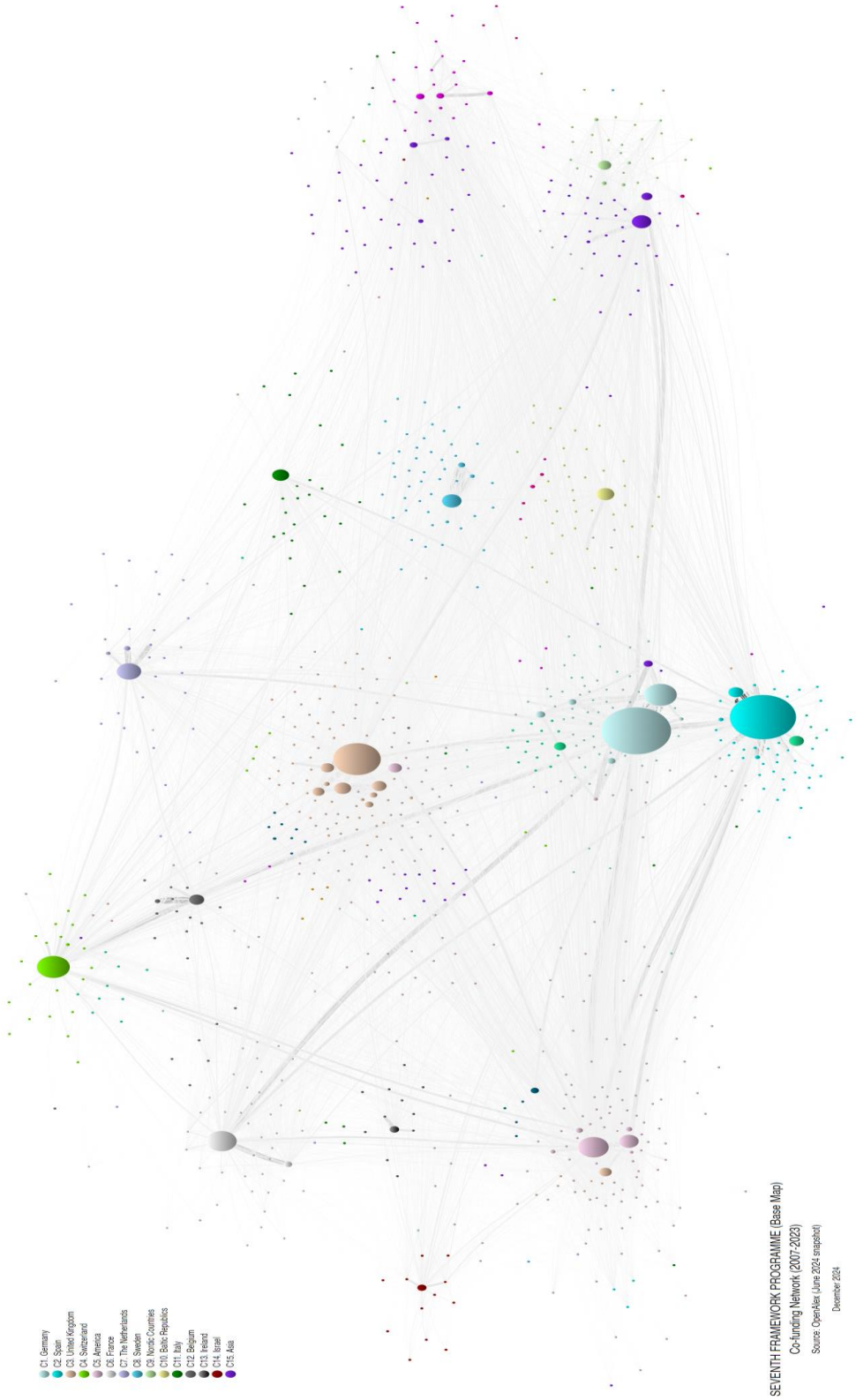


Figure 2. Co-funding base map. Sources: CORDIS and OFR (2007-2023).

Table 3 presents the ten most productive co-funding organisations in FP7. German and Spanish national research foundations have by far the highest number of co-funded projects.

Table 3. Ten most productive international funding agencies in FP7.

Funder	*Country*	*Publications*	*Projects*
German Research Foundation (DFG)	Germany	1,840	1,625
Agencia Estatal de Investigación (AEI)	Spain	1,739	1,899
Engineering and Phys. Sci. Res. Council (EPSRC)	United Kingdom	1,254	988
Government of Germany	Germany	865	619
Schweizerische Nationalfonds (SNSF)	Switzerland	864	765
National Science Foundation (NSF)	United States	803	855
Agence Nationale de la Recherche (ANR)	France	794	684
Dutch Research Council (NWO)	The Netherlands	641	490
FORMAS	Sweden	514	400
National Natural Science Foundation (NSFC)	China	510	776

The data from Table 4 have been used to label clusters based on homogeneity; homogeneity in this case is shown primarily on the basis of geographical links. Each cluster's label indicates the cluster's predominant country or geographical area. There are clusters that are more heterogeneous, like C5, made up of funders from the United States, Canada, Brazil and Chile, C10, which contains funders from Finland and Baltic republics, and C15, which consists of funders from southeast Asia.

Other groups are much more homogeneous. For example, 83.9% of the funders in C11 are Italian. Co-funders from the Netherlands make up 93.9% of C7. Only C9 has a homogeneity of under 50%; Norwegian and Danish agencies account for only 36.4% of the group.

The country proportions in Table 5 reveal that all the funders from the Baltic countries and Ireland fall into C10 and C13, respectively. Other countries, like Sweden (96%), Israel (92.9%) and Spain (91.3%), have practically all their funders in C8, C14 and C2. Asia is an exception: only 39.8% of its funding agencies are members of C15. The proportions of non-European funders are shown in blue. Interestingly, four out of 10 funders are Asian, British or American.

Table 4. Proportion of funders by cluster (nationality).

Country/Region	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*	*C10*	*C11*	*C12*	*C13*	*C14*	*C15*	*Total*
DEU	61.7	0.0	1.1	2.5	1.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0	1.1	6.8
ESP	0.0	71.2	0.5	0.0	2.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.8
GBR	1.1	1.7	42.1	7.5	9.0	0.0	0.0	0.0	1.3	0.0	3.2	0.0	0.0	0.0	1.1	10.4
CHE	2.1	0.0	2.1	57.5	0.8	0.0	0.0	0.0	2.6	0.0	0.0	0.0	0.0	0.0	1.1	3.4
America	9.6	10.2	33.7	2.5	76.7	1.8	3.0	2.0	5.2	6.7	3.2	20.8	29.4	15.0	9.1	22.4
FRA	1.1	1.7	2.6	2.5	1.5	78.2	3.0	0.0	0.0	0.0	3.2	4.2	0.0	0.0	1.1	6.0
NLD	2.1	0.0	1.1	2.5	0.8	1.8	93.9	0.0	1.3	0.0	0.0	8.3	0.0	0.0	0.0	4.3
SWE	0.0	0.0	0.0	0.0	0.8	0.0	0.0	94.1	1.3	0.0	0.0	0.0	0.0	0.0	0.0	5.2
Scandinavia	0.0	0.0	0.5	0.0	0.0	1.8	0.0	0.0	36.4	4.4	3.2	0.0	0.0	0.0	1.1	3.6
Baltic	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	82.2	0.0	0.0	0.0	0.0	0.0	3.9
ITA	1.1	1.7	0.5	0.0	0.8	5.5	0.0	0.0	0.0	2.2	83.9	0.0	0.0	0.0	2.3	3.8
BEL	0.0	0.0	0.5	2.5	0.0	10.9	0.0	0.0	0.0	0.0	0.0	58.3	0.0	0.0	1.1	2.4
IRL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	70.6	0.0	0.0	1.3
ISL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	65.0	1.1	1.5
Asia	2.1	1.7	7.4	2.5	2.3	0.0	0.0	0.0	42.9	4.4	0.0	0.0	0.0	0.0	42.0	9.7
EU-14	16.0	10.2	0.5	20.0	0.0	0.0	0.0	0.0	0.0	0.0	3.2	4.2	0.0	0.0	1.1	3.4
EU-13	3.2	1.7	0.5	0.0	0.0	0.0	0.0	0.0	2.6	0.0	0.0	4.2	0.0	0.0	25.0	3.1
Oceania	0.0	0.0	4.2	0.0	3.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4
Africa	0.0	0.0	2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.6
Rest of Europe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.9	6.5	0.0	0.0	0.0	0.0	15.0	11.4	2.1

Table 5. Proportion of funders by country.

Country/Regio	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*	*C10*	*C11*	*C12*	*C13*	*C14*	*C15*
DEU	89.	0.0	3.1	1.5	3.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.5	1.5
ESP	0.0	91.	2.2	0.0	6.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GBR	1.0	1.0	80.	3.0	12.	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0
CHE	6.1	0.0	12.	69.	3.0	0.0	0.0	0.0	6.1	0.0	0.0	0.0	0.0	0.0	3.0
America	4.2	2.8	29.	0.5	47.	0.5	0.5	0.5	1.9	1.4	0.5	2.3	2.3	1.4	3.7
FRA	1.8	1.8	8.8	1.8	3.5	75.	1.8	0.0	0.0	0.0	1.8	1.8	0.0	0.0	1.8
NLD	4.9	0.0	4.9	2.4	2.4	2.4	75.	0.0	2.4	0.0	0.0	4.9	0.0	0.0	0.0
SWE	0.0	0.0	0.0	0.0	2.0	0.0	0.0	96.	2.0	0.0	0.0	0.0	0.0	0.0	0.0
Scandinavia	0.0	0.0	2.9	0.0	0.0	2.9	0.0	0.0	82.	5.9	2.9	0.0	0.0	0.0	2.9
Baltic	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.	0.0	0.0	0.0	0.0	0.0
ITA	2.8	2.8	2.8	0.0	2.8	8.3	0.0	0.0	0.0	2.8	72.	0.0	0.0	0.0	5.6
BEL	0.0	0.0	4.3	4.3	0.0	26.	0.0	0.0	0.0	0.0	0.0	60.	0.0	0.0	4.3
IRL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.	0.0	0.0
ISL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	92.	7.1
Asia	2.2	1.1	15.	1.1	3.2	0.0	0.0	0.0	35.	2.2	0.0	0.0	0.0	0.0	39.
EU-14	45.	18.	3.0	24.	0.0	0.0	0.0	0.0	0.0	0.0	3.0	3.0	0.0	0.0	3.0
EU-13	10.	3.3	3.3	0.0	0.0	0.0	0.0	0.0	6.7	0.0	0.0	3.3	0.0	0.0	73.
Oceania	0.0	0.0	61.	0.0	38.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Africa	0.0	0.0	83.	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.
Rest of Europe	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.	25.	0.0	0.0	0.0	0.0	15.	50.
TOTAL	9.8	6.2	19.	4.2	13.	5.7	3.4	5.3	8.0	4.7	3.2	2.5	1.8	2.1	9.2

C8 is a special cluster. It contains 96% of the Swedish funding agencies, which in their turn make up 94.1% of the agencies in this cluster. This means the agencies are extremely autonomous or extremely isolated (averse to co-funding with institutions from other countries or regions). On the other hand, 40% of Asian institutions only make up 42% of cluster C15. C15 is the most heterogenous, most dependent cluster, as might be expected of a cluster of non-European funders. The same may be said, although to a lesser degree, about agencies from America and the United Kingdom. In this sense, it should be noted that this study maintained the current EU-27 group, even though the United Kingdom was an EU member country during FP7.

In C1 61.7% are German agencies, accounting for 89.2% of all German funders. EU-14 countries, like Austria (45.5%), have significant weight in this cluster. Something similar happens in C2, where 71.2% of funders are Spanish, accounting in their turn for 91.3% of all Spanish funders. In C3 80% of funders are British, but they account for less than half (42.1%) of the British agencies in the network. Important co-funding with American countries and with practically all the African countries and Oceania can be seen.

Consequently, the most heterogeneous clusters are C5, C9 and C15. In cluster C5 76.6% are American funders, but they make up only 47.7% of the region's funders; this indicates wide scattering. The scattering is even greater in C15, where less than 40% of the funding agencies are from Asia. These anomalies can be explained by the fact that the regions in question are not directly involved in FP7 funding and their collaboration takes place in different clusters obeying diverse interests, where the national or regional effects are less intense.

The base map of FP7 co-funders includes those agencies that are mentioned in the acknowledgements of at least 10 publications. Using this threshold can help augment the effects of regionalisation. In terms of structural indicators (Table 6), it is a large network, with more than 900 nodes. The total number of edges (29,521) is only 6.6% of the possible connections; this indicates low density, although the network is much more dense than other, larger technological networks (Ji et al., 2024).

The average degree (62.35) and degree centralisation (0.56) of the base map are considerably higher than those of the other programmes. This suggests a greater ability to attract funding partners, although the network's centralised structure has few funding agencies in a leading role. The average distance (2.07) is very low, as is the diameter (4), revealing that this is an efficient network with abundant inter-cluster node edges.

Table 6. Structural indicators. Base map, Ideas, People and Health.

Indicators	*Base Map*	*Ideas-ERC*	*People*	*Health*
Nodes	947	431	121	223
Edges	29,521	7,894	2,480	3,314
Density	6.59	8.52	34.16	13.39
Average degree	62.35	36.63	40.99	29.72
Degree centralisation	0.56	0.43	0.49	0.40
Betweenness centralisation	0.085	0.077	0.057	0.080
Average distance	2.07	2.16	1.70	2.06
Diameter	4	4	3	5

In addition to the characterisation of the base map, Table 6 contains the structural indicators of FP7's top three funding programmes by number of publications and projects. FP7-IDEAS-ERC is the programme with the most publications (3,174) and the most projects mentioned in OFR acknowledgements (1,185). FP7-PEOPLE is acknowledged in 564 publications mentioning 360 projects. Lastly, FP7-HEALTH (part of FP7-COOPERATION) is named in 1,055 publications mentioning 308 projects.

The subprogramme with the most funding agencies is IDEAS-ERC (431), and, although it is also the network with the most edges, its density (8.52) is less than that of the other two subprogrammes. This is to be expected, since IDEAS-ERC provides funding for individual researchers. What is noteworthy is the high number of co-funders its grantees attract.

PEOPLE has an extremely high density (34.16) and great compactness, with the lowest betweenness centralisation (0.06), the highest average degree (40.99), the highest degree centralisation (0.43) and the lowest average distance (1.7) of the three subprogrammes.

HEALTH does not stand out in terms of any of its indicators. It is not the most numerous network (223 nodes) or the densest (13.39%). Its diameter is the greatest (5), its betweenness centralisation is the highest of the subprogrammes (0.080), and its degree centralisation is the lowest (0.40). The picture is one of an incohesive, less efficient, more centralised network with a few important nodes to which most of the edges are connected.

Discussion and Conclusions

Structural analysis of the projects and publications of the main programmes of FP7 reveals high co-funding in this macro-programme. Contributions from the funding agencies of European countries and different regions give the programme an extra boost; it is estimated that, for each euro invested, FP7 generated 11 euros of direct and indirect economic effects in the form of innovations, new technologies and products that help meet social challenges and improve the quality of European science systems (European Commission, 2018).

Surprisingly, the publication acknowledgements downloaded from OFR are found to mention 3,020 projects that do not have associated publications reported in CORDIS. This increases the number of FP7-funded projects with scholarly output by 21%. This discovery highlights the great usefulness of analysis based on multiple sources in this and other kinds of studies. In addition, it emphasises the urgent need for the actors involved in all funding flow processes to be responsible and to report and publish accurate, reliable funding acknowledgements in their research results. Research strategies and policies that facilitate access to such data must continue to be implemented, so the data can be analysed properly and the quality and transparency of research can be improved.

It is found that 40% of co-funding agencies are from non-European countries, thus revealing a high level of international cooperation. Asia and America are especially active. This finding is in line with FP7's strategic objectives, which seek to

strengthen competitive international participation in projects as well as in training and mobility actions.

The comparison of co-funding data on the three main FP7 subprogrammes reveals big differences. The IDEAS-ERC co-funder network is the most numerous, has the most publications and includes acknowledgements of more projects. For a programme aimed at individual grantees, it is surprisingly successful at attracting co-funders, doubly so because IDEAS-ERC does not require cross-border associations. This programme bases a good deal of its work on European associations, which enable effective collaboration, thus helping to comply with the scientific strategies mapped out for the programme, focusing on cutting-edge research. Future studies on the visibility of its research results will help arrive at a clear understanding of the excellence it has attained.

The PEOPLE network is the densest, most compact and most decentralised (least influenced by major nodes). Its intense connections speak to the programme's success in reaching its main objective, which is to connect European researchers and institutions through mobility to foster scientific collaboration. PEOPLE's high degree of co-funding activity is aligned with its actions aimed at coordinating scientific collaboration relationships between institutions on the basis of mobility and other instruments oriented toward the lifelong development of researchers' skills and competences. These results agree with those of the ex-post evaluation of the Framework Programme, which states that FP7 helped establish research networks (European Commission, 2018). The degree of excellence the report also mentions has yet to be corroborated by future analyses of published results' visibility.

The results of IDEAS and PEOPLE contrast with those of HEALTH. HEALTH is one of the main thematic areas of COOPERATION, the programme that manages two thirds of FP7's total budget. The objectives of conducting cooperative research in Europe and with other countries through transnational consortiums partnering industry and academia have been partially met from the structural perspective. Although HEALTH has a considerable number of collaborators and moderate density, it has a small number of nodes centralising relationships, and those nodes play too strong a role.

Results by regions show that some clusters are highly independent, while others are dependent. Among the independent clusters, the Swedish funding agencies are extremely isolated from other co-funders. Ninety-four percent of the nodes in cluster C8 are Swedish, and they in their turn account for 96% of all the Swedish funders in the entire network, leaving little margin for international co-funding for the work they sponsor.

Among the dependent clusters, there are two kinds of dependence. First, the dependence of small European countries that establish geography-based ties with larger neighbours, as in the case of Austria (tightly linked to German agencies) and Portugal (tightly linked to Spanish agencies). This sort of dependence reveals collaborations based on social, cultural or linguistic affinities. Another, sharper kind of dependence is found in the agencies of non-European countries, like America and Asia. They appear scattered in diverse clusters, denoting associations that seem to be based more on thematic affinities than on regional or social ties.

Lastly, as stated before, the quantity and quality of the data furnished by FAs are decisive and make a difference in the evaluability of research funding performance. Agencies must set specific mandates for researchers to include clear, precise statements of the funding they have received (which means designing unique project numbers). Researchers have the obligation to acknowledge the support behind their research. Editors must make it easy to report this information, for example, by establishing separate sections where authors must identify their funder and give an unambiguous project number.

Limitations and Future Work

Although this work does not have the disadvantages associated with sample analysis, because it analyses all the publications in Crossref that give an FA and all the projects in CORDIS with FP7 funding, it is not free of limitations. The main drawbacks that limit the scope of the results include poor access to quality funder data, problems in detection and availability of funder award metadata in databases, and errors and omissions in funding information on the part of authors and/or editors.

Furthermore, FA-based evaluation examines only one facet of research work. It fails to explore other aspects of scientific activity, like the number of patents registered, the number of cooperation agreements signed, the number of contracts concluded, young researcher training, conference organisation or scientific equipment procurement or construction.

Also, while the methods, techniques and results presented in this study are extremely helpful for evaluating funding systems, they cannot replace expert judgement in decision making. As editors demand the inclusion of accurate, reliable funding data, readers will trust the results more fully, funders will be able to conduct more accurate analyses of compliance with their objectives and specialists in quantitative studies of science will be able to consolidate this area of study.

Future work to flesh out this analysis should look into the role of funding agencies in highly cited publications, evaluate the influence of co-authorship and co-funding on productivity and publication influence, and analyse the productivity and visibility of the research published in each of the FP7 programmes. Then, quantitative and structural analyses will offer a significant, singular view of compliance with the general objectives of the framework programme and all its subprogrammes.

Acknowledgments

The doctoral dissertation of NSA is funded by Comunidad de Madrid-Spain (ROR: https://ror.org/040scgh75), grant number: PIPF-2022/PH-HUM-25963.

References

Aagaard, K., Mongeon, P., Ramos-Vielba, I., & Thomas, D.A. (2021). Getting to the bottom of research funding: Acknowledging the complexity of funding dynamics. PLoS One, 16(5), e0251488. https://doi.org/10.1371/journal.pone.0251488.

Álvarez-Bornstein, B. (2021). «Acknowledgements» in scientific publications as a tool for analyzing the impact of research funding. [Doctoral dissertation]. Madrid: Universidad Complutense. https://eprints.ucm.es/id/eprint/67595/1/T42836.pdf.

Álvarez-Bornstein, B., & Montesi, M. (2020). Funding acknowledgements in scientifc publications: A literature review. Research Evaluation, 29(4), 469-488. https://doi.org/10.1093/reseval/rvaa038.

Ardanuy, J., Arguimbau, L., Borrego, A. & Sulé, A. (2023). Social Sciences and Humanities research funded under the European Union Seventh Framework Programme (2007-2013): the challenge of retrieving its scholarly outputs [preprint]. 27th International Conference on Science, Technology and Innovation Indicators. Leiden, September 27-29. https://dapp.orvium.io/deposits/643ff48b2271d2fad515761b/view.

Ardanuy, J., Sulé, A. & Borrego, A. (2024). Participación Española en proyectos de investigación en ciencias sociales y humanidades dentro del 7º Programa Marco de la Unión Europea (2007-2013). Revista Española de Documentación Científica, 47(3), e394. https://doi.org/10.3989/redc.2024.3.1557.

Batagelj, V., Mrvar, A. (2004). Pajek: Analysis and Visualization of Large Networks. In: Jünger, M., Mutzel, P. (eds). Graph Drawing Software. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-18638-7_4.

Blondel, V.D., Guillaume, J.L., Lambiotte, R. Lefebvre E. (2008). Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment, 10, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

Boyack, K.W. (2009). Linking grants to articles: Characterization of NIH grant information indexed in Medline. Proceedings of ISSI, 730-741. https://www.issi-society.org/proceedings/issi_2009/ISSI2009-proc-vol2_Aug2009_batch1-paper-22.pdf.

Costas, R., & van Leeuwen, T.N. (2012). Approaching the "reward triangle": General analysis of the presence of funding acknowledgments and "peer interactive communication" in scientific publications. Journal of the American Society for Information Science and Technology, 63(8), 1647-1661. https://doi.org/10.1002/asi.22692.

European Comission (2007). FP7 in Brief. How to get involved in the EU 7th Framework Programme for Research. A pocket guide to newcomers. European Sources Online. ISBN: 92-79-04805-0. https://www.europeansources.info/record/fp7-in-brief-how-to-get-involved-in-the-eu-7th-framework-programme-for-research-a-pocket-guide-to-newcomers/.

European Commission (2016). Commission presents its evaluation of the FP7 Framework Programme for Research. Brussels: European Comission. https://ec.europa.eu/commission/presscorner/detail/en/MEMO_16_146.

European Comission (2018). Commitment and coherence. Ex post evaluation of the 7th EU Framework Programme (2007-2013). https://op.europa.eu/es/publication-detail/-/publication/7e74df87-ebb0-11e8-b690-01aa75ed71a1.

Grassano, N., Rotolo, D., Hutton, J., Lang, F., & Hopkins, M.M. (2017). Funding data from publication acknowledgments: Coverage, uses, and limitations. Journal of the Association for Information Science and Technology, 68(4), 999-1017. https://doi.org/10.1002/asi.23737.

Jin,Y.; Cao, X.; Ma, H. (2024). Evolution and characteristics of Crossover Innovation Network of Emerging Technologies: a study based on patent data of the self-driving car technology. Transinformação, 36, e247316. https://doi.org/10.1590/2318-0889202436e247316.

Kamada, T., Kawai, S. (1989). An algorithm for drawing general undirected graphs. Information Processing Letters, 31(1), p. 7-15. https://doi.org/10.1016/0020-0190(89)90102-6.

Kramer, B., & de Jonge, H. (2022). The availability and completeness of open funder metadata: Case study for publications funded by the Dutch Research Council. Quantitative Science Studies, 3(3), 583-599. https://doi.org/10.1162/qss_a_00210.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology, 60(2), p. 348-362. https://doi.org/10.1002/asi.20967.

Mugabushaka AM. (2020). Linking Publications to funding at project level: a curated dataset of publications reported by FP7 projects. arXiv. DOI: https://doi.org/10.48550/arXiv.2011.07880.

Mugabushaka, A.M., van Eck, N.J., & Waltman, L. (2022). Funding COVID-19 research: insights from an exploratory analysis using open data infrastructures. Quantitative Science Studies, 3(3), 560-582. https://doi.org/10.1162/qss_a_00212.

Perianes-Rodríguez, A, & Ruiz-Castillo, J. (2015). Multiplicative versus fractional counting methods for co-authored publications. The case of the 500 universities in the Leiden Ranking. Journal of Informetrics, 9(4), p. 974-89. https://doi.org/10.1016/j.joi.2015.10.002.

Perianes-Rodríguez, A., Waltman, L., & Van Eck, N.J. (2016). Constructing bibliometric networks: A comparison between full and fractional counting. Journal of Informetrics,10(4), 1178-1195. https://doi.org/10.1016/j.joi.2016.10.006.

Perianes-Rodríguez, A., Olmeda-Gómez, C., Delbianco, N.R., & Cabrini, M.C. (2024a). Public funding accountability: A linked open data-based methodology for analysing the scientific productivity and influence of funded projects. Scientometrics. https://doi.org/10.1007/s11192-024-04975-8.

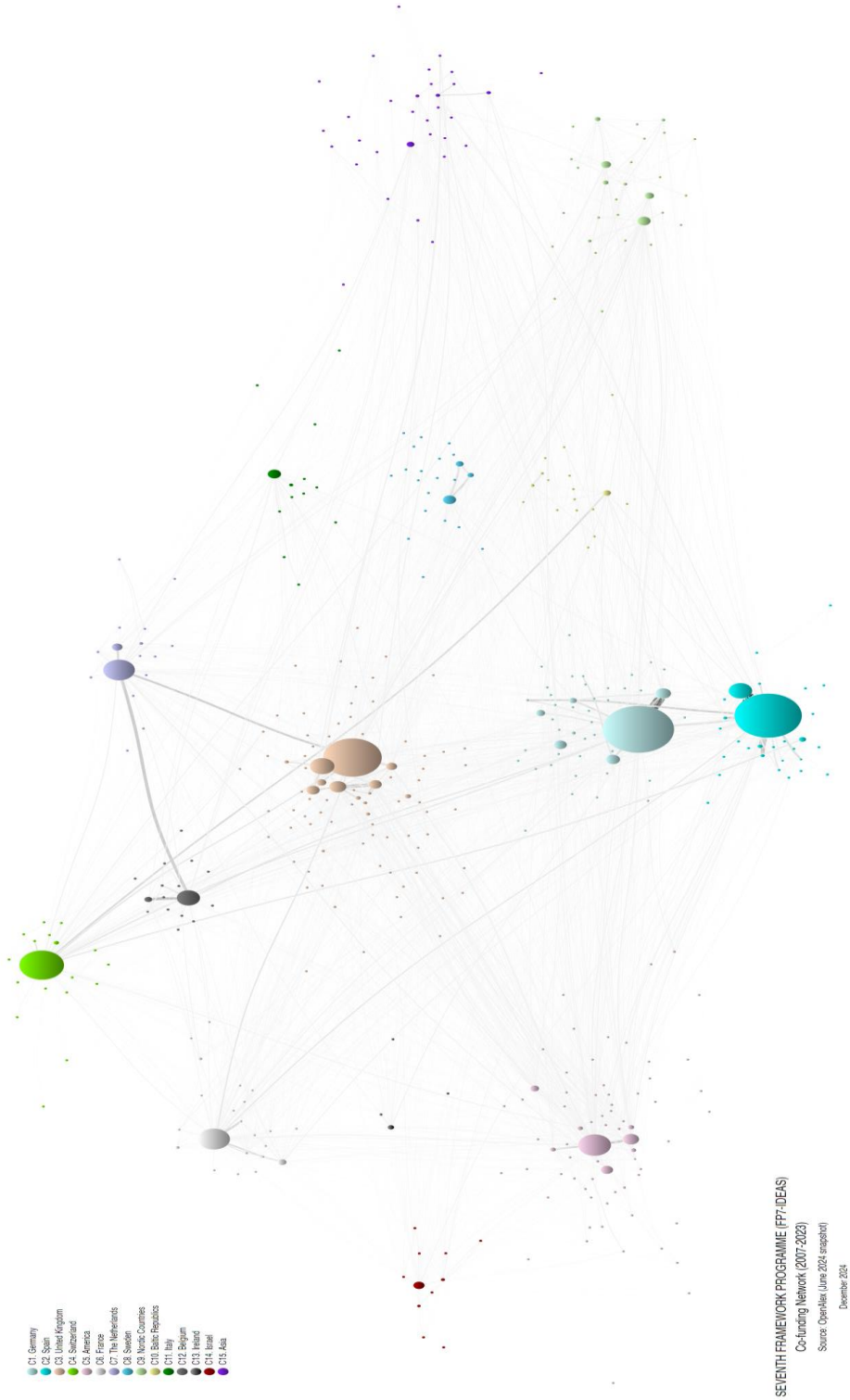
Perianes-Rodriguez, A., & Silva-Alés, N. (2024b). Co-funding networks as a new tool in research evaluation: a linked open data-based study of the Seventh Framework Programme projects. Dataset [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14502483.

ROR. (2024). What is ROR? [Software]. https://ror.org/.

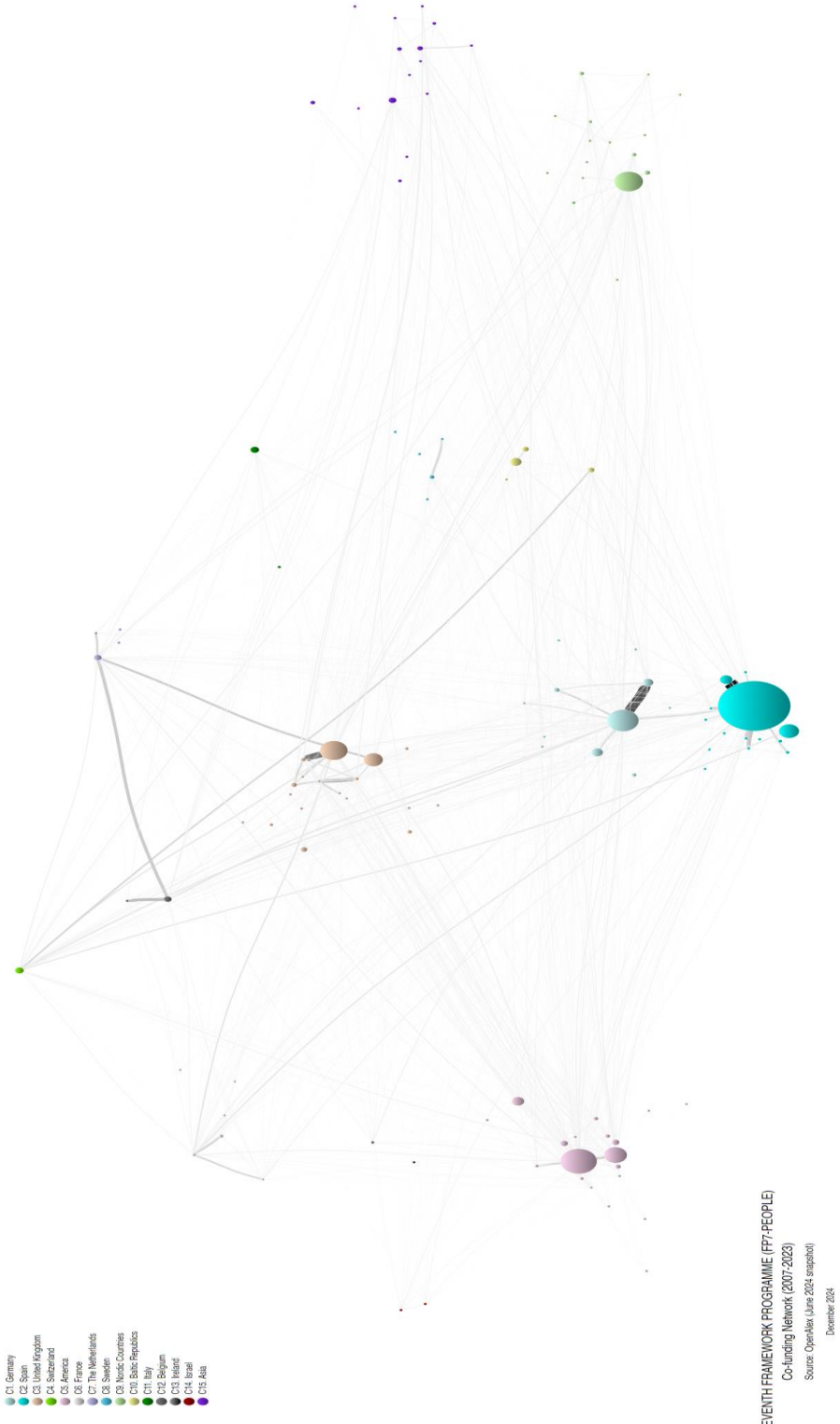
Waltman, L., & Van Eck, N.J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. Journal of Informetrics, 9(4), 872–894. https://www.sciencedirect.com/science/article/pii/S1751157715300456.

Wang, J. & Shapira, P. (2011) Funding acknowledgement analysis: an enhanced tool to investigate research sponsorship impacts: the case of nanotechnology. Scientometrics 87, 563-586. https://doi.org/10.1007/s11192-011-0362-5.

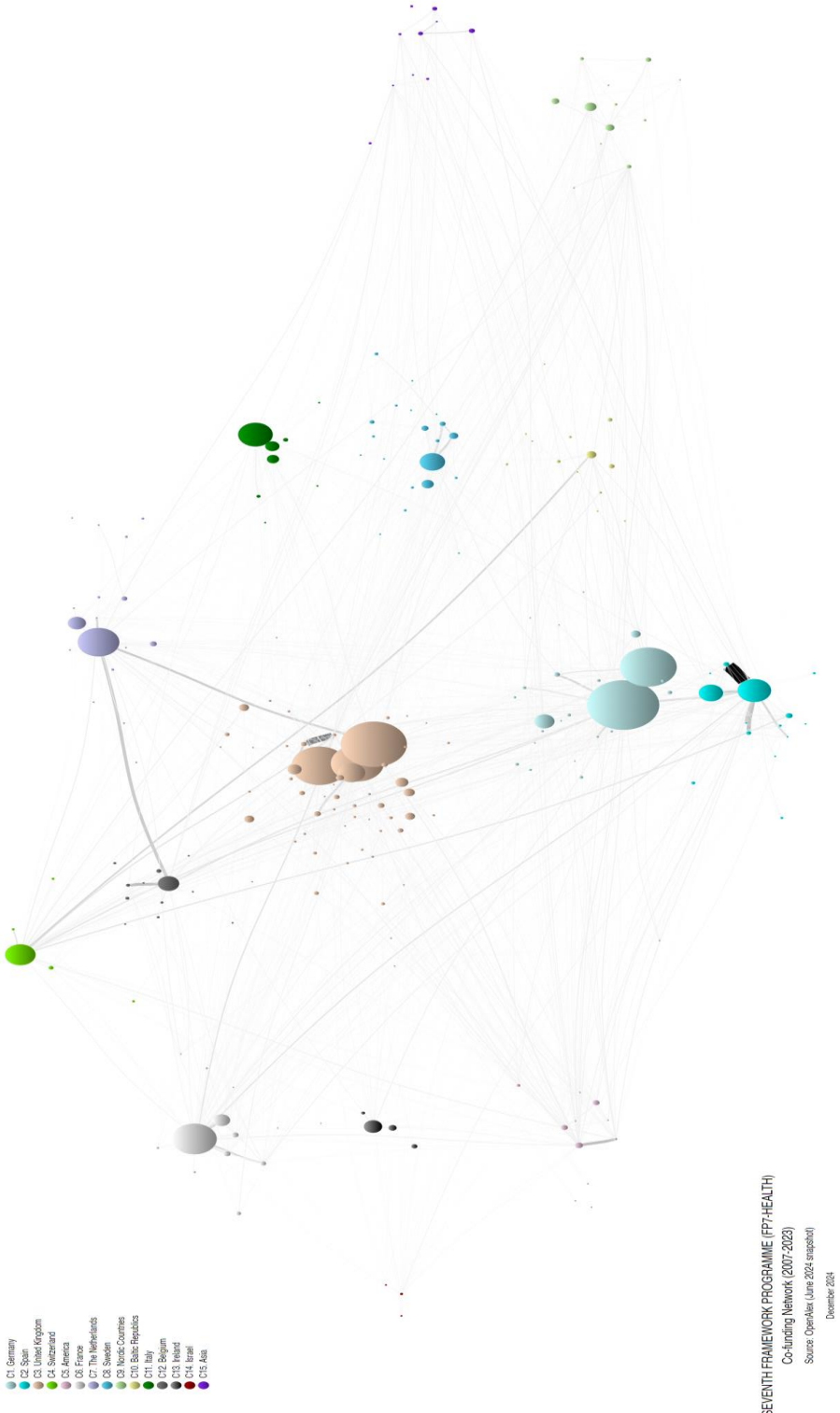
Annex 1. Co-funding Map. FP7-IDEAS-ERC. Sources: CORDIS and OFR (2007-2023).



Annex 2. Co-funding Map. FP7-PEOPLE. Sources: CORDIS and OFR (2007-2023).



Annex 3. Co-funding Map. FP7-HEALTH. Sources: CORDIS and OFR (2007-2023).



https://doi.org/10.51408/issi2025_164

Scientific Landscape in the South Caucasus: A Comparative Analysis of Armenia, Azerbaijan, and Georgia (2012–2024)

Edita Gzoyan[1], Aram Mirzoyan[2], Gevorg Kesoyan[3], Mariam Yeghikyan[4], Simon Hunanyan[5], Shushanik Sargsyan[6]

[1]*editagzoyan@gmail.com,* [2]*aram.mirzoyan@asnet.am,* [3]*gevorgkesoyaned@gmail.com,*
[4]*mariam_yeghikian@mail.ru,* [5]*simhunanyan@gmail.com,* [6]*shushaniksargsyan8@gmail.com*
Institute for Informatics and Automation Problems of NAS RA, 1 Paruyr Sevak St, 0014,
Yerevan (Republic of Armenia)

Abstract

This article presents a comparative analysis of the scientific output of Armenia, Azerbaijan, and Georgia over the period from 2012 to 2024. Using data from the Web of Science international database, the study will examine the research productivity and impact of these countries, highlighting trends, policies, and developments that have influenced their scientific landscapes. Special attention will be given to journal indexing policies, particularly those related to the inclusion of national and local journals in the Web of Science (WoS) and their impact on the number of publications form the perspective states.

The analysis begins by situating the scientific efforts of these republics within their historical context, reflecting on their roles around 33 years after regaining independence from the Soviet Union. It then focuses on the post-independence period, with a particular emphasis on the past decade. The article evaluates key indicators such as publication volume, citation metrics, and international collaborations. Special attention is given to recent policies and strategies implemented in Armenia, Azerbaijan, and Georgia to foster research and development, and their outcomes in terms of scientific progress.

This study aims to provide a comprehensive understanding of the similarities and differences in the scientific trajectories of these nations and their positions in the global scientific community during the specified timeframe of 2012-2024.

Introduction

The South Caucasus region, encompassing Armenia, Azerbaijan, and Georgia, has a complex scientific landscape shaped by historical legacies, political developments, and economic transformations. During the Soviet era, these three republics played distinct yet interconnected roles in the USSR's centralized scientific system. Research institutions, academies of sciences, and universities in the region benefited from substantial state funding and integration into the broader Soviet knowledge production framework. However, the dissolution of the Soviet Union in 1991 led to a period of economic and institutional decline, significantly affecting the scientific and technological capacities of these newly independent states (Chankseliani et all.2018; Chankseliani et. all, 2021).

Over the past three decades, Armenia, Azerbaijan, and Georgia have pursued different paths in revitalizing their research sectors, influenced by national policies, international collaborations, and economic constraints. Armenia has increasingly positioned itself as a hub for information technology and innovation, leveraging its strong diaspora connections and historical scientific expertise (Abramo et al., 2025; Gzoyan et al., 2023). Azerbaijan, with its resource-rich economy, has prioritized

applied research in energy and technology, seeking to integrate scientific advancements into its economic diversification efforts (Humbatova, 2021). Georgia, meanwhile, has focused on strengthening ties with European research institutions, aiming to modernize its academic infrastructure and increase participation in international projects (Chagelishvili, 2025).

Scientometric analysis provides a valuable tool for understanding the evolution and impact of research output in these countries. By examining publication trends, citation metrics, and international collaborations, this study aims to assess the scientific performance of Armenia, Azerbaijan, and Georgia within the broader global and regional contexts. Through this approach, we seek to identify key trends, challenges, and opportunities that shape the research landscapes of these nations, contributing to a deeper understanding of their scientific trajectories in the post-Soviet era.

This research builds upon the findings of our study "Comparative Analysis of the Scientific Output of Armenia, Azerbaijan, and Georgia," which examined the research productivity and collaboration patterns of these three South Caucasus countries up until 2013 (Gzoyan et al., 2015). While that study provided a foundational understanding of the region's scientific landscape, significant developments have occurred over the past decade, necessitating an updated analysis. Since 2013, Armenia, Azerbaijan, and Georgia have implemented various policy reforms, expanded international collaborations, and witnessed shifts in research funding and institutional priorities. This study aims to assess these changes by employing a scientometric approach to evaluate publication trends, citation impact, and regional as well as global integration in scientific research. By analyzing the latest data, this research seeks to provide a comprehensive overview of the evolving scientific output in the South Caucasus, identifying both progress and persistent challenges in the region's research landscape.

After seventy years of Soviet rule, Armenia, Azerbaijan and Georgia—along with twelve other post-Soviet states—regained their independence in 1991. However, independence came with significant political, social, and economic challenges, including crucial decisions regarding regional and global integration. In an effort to maintain its influence over the former Soviet republics, Russia initiated a new "integration project" almost simultaneously with the dissolution of the USSR—the Commonwealth of Independent States (CIS), also referred to as the Russian Commonwealth. All former Soviet republics, except for the Baltic states (Latvia, Lithuania, and Estonia), joined the CIS, which, as some analysts have described, functioned as a "civilized divorce" between Russia and its former republics. However, over time, the CIS proved to be largely ineffective, with its political relevance steadily declining.

Recognizing the limitations of the CIS, Russia launched new reintegration initiatives aimed at consolidating its influence in the post-Soviet space. These include the Collective Security Treaty Organization (CSTO), a military alliance formed to enhance regional security, and the Eurasian Economic Union (EAEU), modeled to some extent after the European Union's economic integration framework. Armenia, seeking to balance its foreign policy between East and West, initially engaged in

cooperation with both Russia-led and EU-led initiatives. The country's relations with the European Union began in 1999 with the signing of the EU-Armenia Partnership and Cooperation Agreement, which facilitated collaboration in political dialogue, economic development, trade, democracy, human rights, law-making, and cultural exchange. Armenia further participated in two key EU programs: the European Neighborhood Policy (ENP), since 2004, and the Eastern Partnership (EaP), since 2009, strengthening its engagement with European institutions. However, in 2013, Armenia opted to join the Eurasian Economic Union (EAEU) instead of signing an Association Agreement with the EU, marking a shift in its geopolitical trajectory (Sargsyan et al., 2020).

Georgia and Azerbaijan, while also part of the CIS in the 1990s, took divergent paths in their post-Soviet integration strategies. Georgia, following the 2003 Rose Revolution, actively pursued a pro-Western foreign policy, prioritizing deeper integration with the European Union and NATO. The country formally left the CIS in 2008 after its war with Russia and later signed an Association Agreement with the EU in 2014, reinforcing its European aspirations.

Meanwhile, Azerbaijan, despite being a CIS member, adopted a more independent and pragmatic foreign policy, leveraging its vast energy resources to maintain strategic partnerships with both Russia and Western countries. While Azerbaijan has engaged in select EU initiatives, such as the Eastern Partnership, it has refrained from deeper political or economic integration with either the EU or Russia-led blocs, preferring a non-aligned approach that maximizes its geopolitical leverage. Azerbaijan, however, has fostered exceptionally close relations with Turkey, a partnership rooted in deep historical, cultural, and linguistic ties (Mikail, et. al. 2019). The two countries often emphasize their bond through the phrase "One Nation, Two States," reflecting their strategic alliance in political, economic, and military domains. Turkey has played a crucial role in Azerbaijan's military modernization, particularly evident in the 2020 Nagorno-Karabakh war, where Turkish military support, including drone technology, significantly influenced the conflict's outcome. Economically, Turkey is a key transit country for Azerbaijani oil and gas exports, especially through major energy projects such as the Baku-Tbilisi-Ceyhan (BTC) pipeline and the Trans-Anatolian Natural Gas Pipeline (TANAP). These strong ties also extend to scientific and technological cooperation, with both countries engaging in joint research projects, educational exchange programs, and innovation initiatives, particularly in defense and energy sectors.

Since the early 1990s Turkey (Türkiye), after half a century of political, economic and cultural estrangement, has resumed its multifaceted engagement with the Caucasus (Sukiasyan et. al., 2025), alongside post-Soviet Central Asia and the post-communist Balkans. In 1992 Turkey became a founder of the Black Sea Economic Cooperation Organisation, with its headquarters in Istanbul; and during the 2000s and the 2010s asserted itself as a key player in major regional energy and transportation projects, including the Baku-Tbilisi-Ceyhan (BTC) and the Baku-Tbilisi-Erzerum (BTE) gas pipelines, the Trans-Anatolian Natural Gas Pipeline (TANAP), and the Trans-Caspian East-West Middle Corridor (known as the “Middle

Corridor") initiative connecting China and Turkey via Turkic Central Asia, Azerbaijan and Georgia (Yemelianova, 2023).

These divergent integration paths among Armenia, Georgia, and Azerbaijan have had direct implications for their scientific and technological development. While Georgia has sought closer collaboration with European research institutions, Armenia has attempted to balance cooperation between Russian and Western scientific networks. Azerbaijan, with its resource-driven economy and close ties with Turkey, has primarily invested in applied research tied to energy, infrastructure, and defense technology. This study examines how these geopolitical choices have influenced the scientific output of the three South Caucasus nations, analyzing recent trends, collaborations, and the broader regional research landscape.

Methods

This study is based on data retrieved from the Web of Science (WoS), InCites, and Journal Citation Report (JCR). The analysis encompasses scholarly publications affiliated with Armenia, Azerbaijan, and Georgia, indexed in WoS during the period 2012–2024. Citation data for the same timeframe were also included. The document types analyzed comprise WOS all types of documents.

Consistent with the challenges identified by Glänzel and Schlemmer (2009), accurately retrieving publications by country affiliation during the Soviet era presented methodological difficulties due to inconsistencies in institutional naming and geopolitical classifications. To address this, comprehensive search strategies were employed using both official and variant country names. In the case of Armenia, entries erroneously indexed under *Armenia (Colombia)* were manually identified and excluded. Similarly, for Georgia, records associated with the U.S. state of Georgia were filtered out to ensure the accuracy of national attribution.

To identify national/local journals, the Journal Citation Reports (JCR) were consulted. Additionally, a targeted search was conducted for journal titles containing the keywords "Armenian," "Georgian," and "Azerbaijani" to capture region-specific scholarly output not readily identifiable through affiliation data alone.

Analysis of Publication Trends in the South Caucasus (2012–2024)

Figure 1 illustrates the yearly distribution of scientific publications from Armenia, Georgia, and Azerbaijan as indexed in the Web of Science (WoS) between 2012 and 2024. The data reveal several notable trends and divergences in the scientific output of these three South Caucasus countries.

From 2012 to around 2018, all three countries displayed relatively modest but steady growth in publication output, with Armenia maintaining a slight lead over its neighbors. During this period, Armenia and Georgia showed gradual increases, while Azerbaijan's output remained close to theirs but slightly more variable.

A noticeable shift occurs around 2019–2020, where Azerbaijan's publication count begins to accelerate more rapidly, surpassing both Armenia and Georgia. This upward trajectory becomes particularly sharp between 2022 and 2024, culminating in a dramatic rise in 2024 where Azerbaijan reaches over 3,500 publications—nearly double that of Armenia and significantly more than Georgia.

Armenia and Georgia also experienced growth during this period, though at a more moderate pace. By 2024, Armenia surpassed 1,800 publications, while Georgia approached 1,600. The overall upward trend for all three countries suggests growing engagement in international research and increased visibility in indexed journals, but Azerbaijan's particularly steep rise in recent years indicates a potentially significant shift in research funding, institutional strategies, or international collaboration efforts.

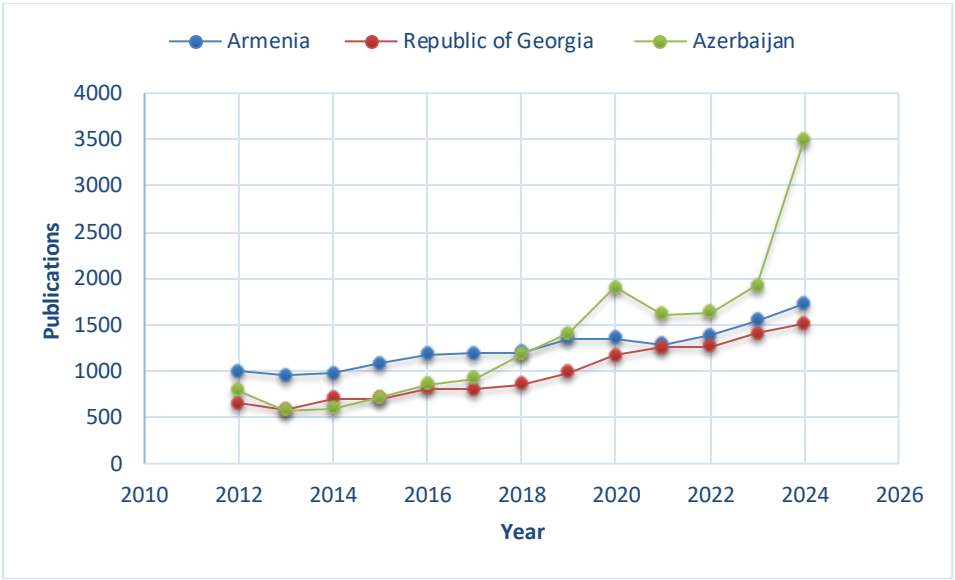


Figure 1. Yearly distribution of publications from Armenia, Georgia, and Azerbaijan (WoS, 2012–2024).

The next **Figure 2** shows the number of citations received by publications from Armenia, Georgia, and Azerbaijan between 2012 and 2024. Armenia had high citation numbers in the early years, especially in 2012 and 2018, reaching around 30,000 and 39,000 citations, respectively. Georgia saw a major peak in 2014, also with nearly 30,000 citations, while Azerbaijan showed more gradual growth, peaking in 2020 with close to 28,000 citations.

From 2012 to 2018, Armenia consistently had the most citations. Starting in 2019, Azerbaijan caught up and became the leading country in 2020. In the later years (2021–2024), citation numbers dropped for all three countries, but Azerbaijan maintained a relative lead, especially in 2023 and 2024.

Armenia's sharp rise in 2018 and Georgia's spike in 2014 stand out from the general patterns. Azerbaijan's peak in 2020 might be linked to increased international collaboration or activity in highly cited research fields.

Citation numbers declined significantly for all three countries after 2020. Armenia was most affected, with fewer than 5,000 citations by 2024. Possible reasons include a natural citation delay for recent publications, reduced research output, or external factors such as the COVID-19 pandemic.

The data reflect changing scientific visibility in the South Caucasus region. Armenia was the leader in earlier years, but Azerbaijan gained ground in recent times. Georgia's performance was more uneven, with one standout year in 2014. The overall decline in citations after 2020 suggests broader challenges that may require national-level strategies to improve research impact and visibility.

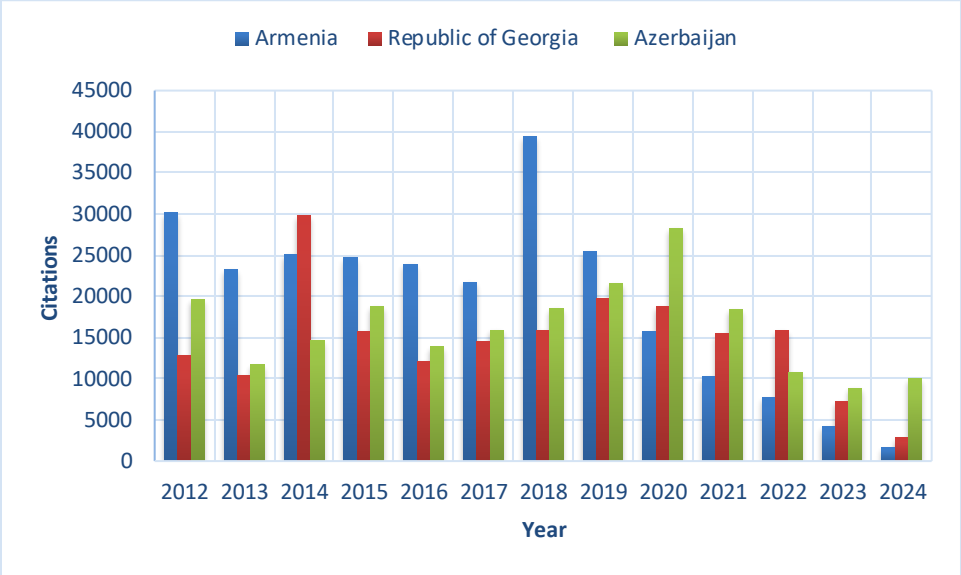


Figure 2. Annual distribution of citations received by publications from Armenia, Georgia, and Azerbaijan (WoS, 2012–2024).

The next focus of the research was on international collaboration of the tree states. The data in **Table 1** highlights the leading international partners in scientific collaboration for Armenia, Georgia, and Azerbaijan, reflecting both geopolitical alignments and strategic research ties.

The United States emerges as the top collaborator for both Armenia and Georgia, underscoring strong academic and institutional connections with North America. This trend aligns with broader political and educational exchanges between these countries and the U.S., as well as the influence of diaspora networks, particularly in Armenia's case.

For Azerbaijan, Turkey ranks first—a reflection of close historical, cultural, and political ties, including extensive bilateral cooperation in higher education and scientific exchange.

Russia ranks second for both Armenia and Azerbaijan, and is notably absent from Georgia's top five. This likely reflects the more strained post-Soviet relationship between Georgia and Russia, especially after the 2008 conflict. In contrast, Armenia and Azerbaijan maintain strong educational and research connections with Russian institutions, rooted in shared language, legacy networks, and continued participation in regional alliances.

Germany and Italy appear consistently across all three countries, suggesting a broader pan-European scientific engagement in the South Caucasus. Germany, in

particular, ranks within the top three for all, highlighting its role as a significant science and innovation partner in the region. France also features prominently for Armenia and Georgia, indicating active bilateral academic initiatives and EU-funded collaborations.

Interestingly, China appears only in Azerbaijan's list (3rd place), pointing to Baku's growing scientific and strategic cooperation with Beijing, likely tied to broader infrastructural and technological investments as part of China's Belt and Road Initiative.

Overall, this table illustrates how geopolitical orientation, historical ties, and strategic interests shape patterns of international research collaboration in the South Caucasus. Armenia and Georgia show stronger alignment with Western institutions, while Azerbaijan maintains close ties with Turkey and is diversifying eastward.

Table 1. Top 5 collaborating countries.

Rank	Armenia	Republic of Georgia	Azerbaijan
1	USA	USA	Turkey
2	Russia	Germany	Russia
3	Germany	United Kingdom	China
4	France	Italy	USA
5	Italy	France	Italy

The Contribution of National/Local Journals to Overall Scholarly Output

As a next step, we have tried to identify the role and share of national/local journals in the number of publications of three republics (Moed et. al., 2021). We have first identified the national journals indexed in the WoS/Scopus (**Table 2**).

Table 2. Local/national journals of Armenia, Georgia and Azerbaijan indexed in WoS (JCR 2023).

Armenia			
№	Name	Categories	Year Entered WoS
1	Armenian Journal of Mathematics	Mathematics	2020
2	New Armenian Medical Journal	Medicine, General & Internal	2020
3	Journal of Contemporary Physics-Armenian Academy of Sciences*	Physics, Multidisciplinary	2010
4	Journal of Contemporary Mathematical Analysis-Armenian Academy of Sciences*	Mathematics	2010
5	Astrophysics	Astronomy & Astrophysics	2004

Republic of Georgia			
№	Name	Category	Year Entered WoS
1	Journal of Homotopy and Related Structures	Mathematics	2010
2	Tbilisi Mathematical Journal	Mathematics	2020
3	Advanced Studies-Euro-Tbilisi Mathematical Journal	Mathematics	2022
4	Memoirs on Differential Equations and Mathematical Physics	Mathematics, Applied	2020
5	Transactions of A Razmadze Mathematical Institute	Mathematics	2020
6	European Journal of Transformation Studies	Political Science	2020
7	Georgian Mathematical Journal*	Mathematics	2009
Azerbaijan			
№	Name	Category	Year Entered WoS
1	Applied and Computational Mathematics	Mathematics, Applied	2009
2	TWMS Journal of Pure and Applied Mathematics	Mathematics; Mathematics, Appliied	2020
3	Proceedings of the Institute of Mathematics and Mechanics	Mathematics	2020
4	Azerbaijan Journal of Mathematics	Mathematics	2020
5	New Materials Compounds and Applications	Chemistry Multidisciplinary; Material Science, Multidisciplinary	2022
6	Processes of Petrochemistry and Oil Refining	Engineering, Chemical	2020
7	Khazar Journal of Humanities and Social Sciences	Social Sciences, Interdisciplinary	2020
8	SOCAR Proceedings	Engineering, Petroleum	2020

An examination of the national journals from Armenia, Georgia, and Azerbaijan indexed in the Web of Science (JCR 2023) reveals notable differences in scale, timing, and disciplinary focus, reflecting broader patterns in national research policy and academic development across the South Caucasus.

As of 2023, Azerbaijan leads the region with eight WoS-indexed journals, followed by Georgia with seven, and Armenia with five. While the numbers may appear

modest in absolute terms, they are significant for understanding each country's strategy for achieving international scientific visibility through academic publishing. A closer look at the chronology of indexing shows a clear regional trend: a major wave of journal inclusion occurred around 2020, likely the result of deliberate national efforts to meet international editorial and peer-review standards. In Armenia, both the *Armenian Journal of Mathematics* and the *New Armenian Medical Journal* were indexed in 2020, adding to earlier entries such as *Astrophysics* (2004) and two journals affiliated with the Armenian Academy of Sciences (2010). Georgia experienced a similar pattern, with four journals added in 2020, though its earliest inclusion, the *Georgian Mathematical Journal*, dates back to 2009. Azerbaijan also saw the majority of its journals indexed in or after 2020, with the exception of *Applied and Computational Mathematics* (2009), and the more recent *New Materials Compounds and Applications* in 2022.

In terms of scientific fields, a heavy concentration in mathematics is evident across all three countries. Armenia's indexed journals are predominantly in mathematics and physics, with a single title in medicine. Georgia's representation is also math-heavy, accounting for five out of seven journals, but it extends modestly into applied mathematics and political science. Azerbaijan, in contrast, demonstrates the broadest disciplinary range, with journals not only in mathematics but also in chemistry, materials science, petroleum engineering, chemical engineering, and interdisciplinary social sciences. This diversity suggests a more deliberate and multifaceted national strategy aimed at integrating a wider spectrum of disciplines into the international scholarly community.

The presence of legacy journals—such as those affiliated with national academies—indicates the role of traditional academic institutions in maintaining continuity, but the recent indexing of newer journals may reflect efforts to modernize editorial practices and increase impact metrics.

This comparison highlights the varied levels of institutional capacity, policy commitment, and strategic direction among the three countries. Azerbaijan appears to be the most proactive in expanding the scope of its internationally recognized journals, while Georgia is steadily reinforcing its strength in the mathematical sciences. Armenia, despite having fewer indexed journals, maintains a strong reputation in foundational sciences, though its narrower disciplinary scope may limit broader academic visibility.

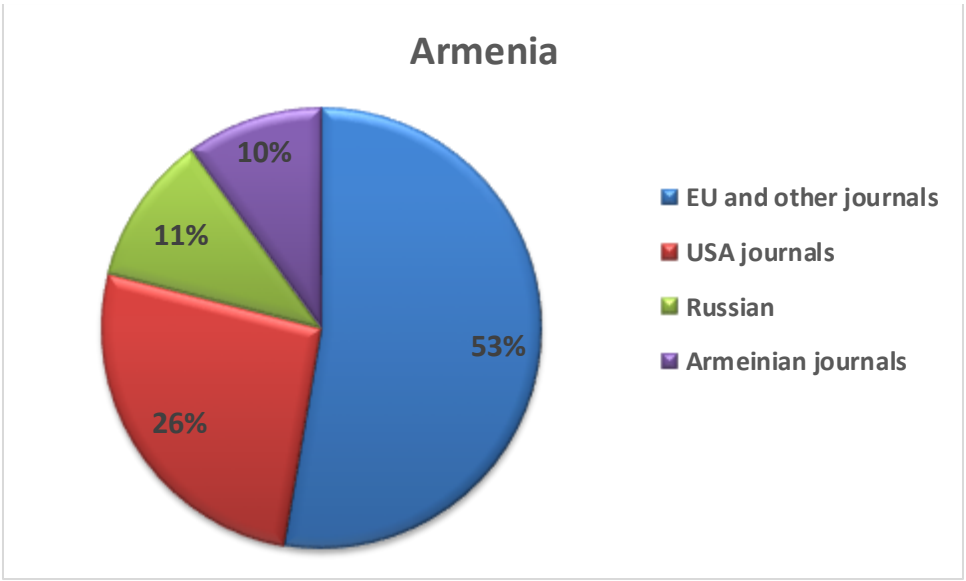
The distribution of publications in foreign journals presents the following picture (**Figure 3**). An analysis of publication data from Armenia, Georgia, and Azerbaijan between 2012 and 2024 reveals distinct trends in the use of Russian, U.S., Turkish, and nationally indexed journals, reflecting varying degrees of geopolitical orientation, linguistic affiliation, and academic strategy. Russian-indexed journals played a prominent role in the publication profile of Armenia and Azerbaijan, but far less so for Georgia.

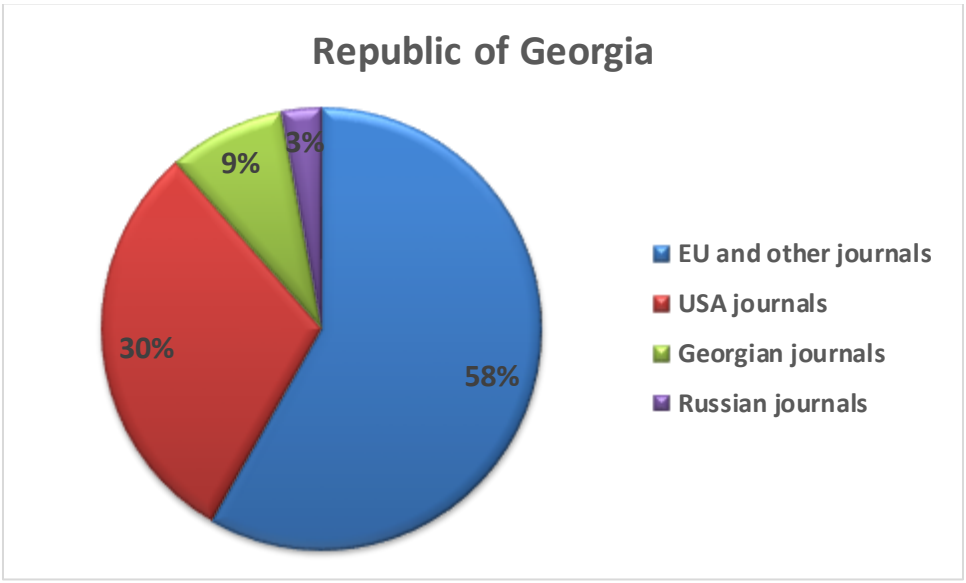
Specifically, 14% of Azerbaijani's publications appeared in Russian journals—by far the highest among the three countries—demonstrating Azerbaijani's strong post-Soviet scholarly ties and the continued use of the Russian language in certain scientific fields. Armenia followed with 3%, also indicating sustained academic

linkage with Russia. In contrast, Georgia's output in Russian journals was relatively marginal, at just 3%, consistent with its broader efforts to pivot toward Western academic integration.

U.S.-indexed journals constituted a significant share of publications across all three countries, but especially in Georgia, where they represented 30% of the total output. This was followed by Armenia at 26% and Azerbaijan at 15%. These figures suggest that all three states are engaged in global scholarly communication, though Georgia leads in Western journal dissemination. Turkish-indexed journals featured in Azerbaijani academic output although surprisingly slightly (2%), underscoring linguistic and cultural proximity as well as growing institutional cooperation between the two countries.

National journals accounted for a considerable share of total outputs, pointing to almost the same share (Armenia and Azerbaijan 10% and Georgia 9%). This representation suggests that for international visibility and citation impact, researchers across the region tend to favor publishing in national journals indexed in WoS and/or Scopus.





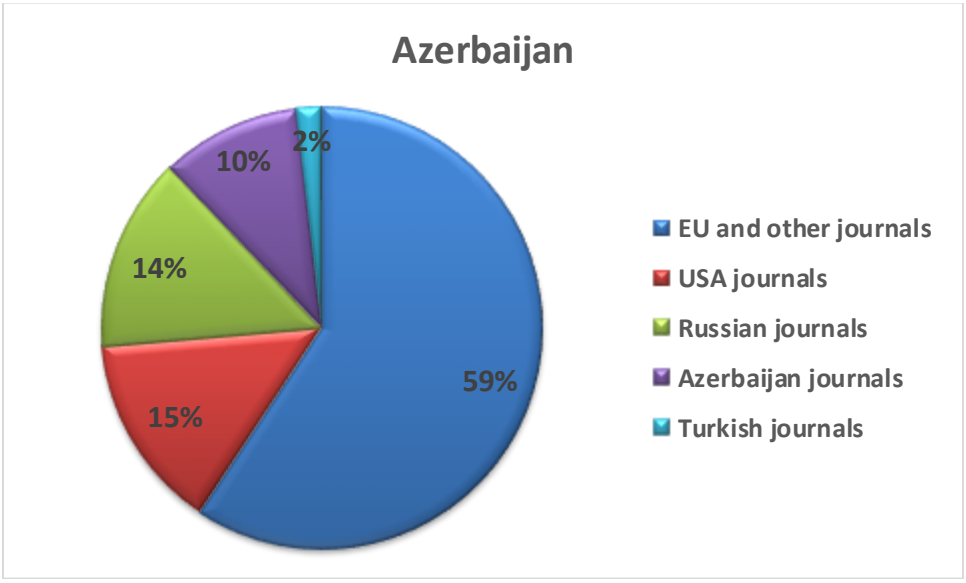


Figure 3. Percentage distribution of publications from Armenia, Georgia, and Azerbaijan by the country of indexed journals (WoS, InCites, 2012–2014), indicating the publishing countries of the respective articles.

Conclusion

This study provides a comprehensive scientometric analysis of the research output of Armenia, Azerbaijan, and Georgia from 2012 to 2024, revealing both shared challenges and divergent trajectories shaped by each country's geopolitical choices, policy priorities, and institutional capacities. While all three nations have demonstrated growth in publication volume and international collaboration, Azerbaijan has experienced a particularly sharp increase in research output since

2020. likely reflecting expanded state investment and broader international engagement, including with China and Turkey.

The data underscore the continued significance of historical and linguistic ties, with Russia remaining a key partner for Armenia and Azerbaijan, but largely absent in Georgia's scientific collaboration. Conversely, Georgia has shown the strongest integration with Western academic institutions, particularly through high publication rates in U.S.-indexed journals and consistent cooperation with European partners. Armenia remains more balanced in its orientation, maintaining ties with both Russian and Western networks.

National and local journals play a growing role in regional research visibility. Azerbaijan leads in the number and disciplinary diversity of WoS-indexed journals, while Georgia excels in mathematics-focused titles. Armenia, although maintaining a strong base in fundamental sciences, appears more limited in scope. The inclusion of national journals in global databases reflects a strategic effort to increase scientific visibility and foster domestic publication ecosystems.

The analysis also highlights notable disparities in citation performance, with Armenia leading in earlier years but Azerbaijan gaining prominence in recent times. The overall decline in citations after 2020 across all three countries may reflect broader structural challenges, such as delays in citation accumulation or disruptions due to the COVID-19 pandemic.

Ultimately, this study demonstrates that while Armenia, Azerbaijan, and Georgia have all made measurable progress in expanding their scientific output, their development paths remain shaped by differing political alignments, economic priorities, and integration strategies. Continued investment in research infrastructure, international collaboration, and journal development will be essential for sustaining and enhancing their positions in the global scientific landscape.

Funding

This work is supported by the Yervant Terzian Armenian National Science and Education Fund (ANSEF) Grant № 25AN:HU-soc-3291.

References

Abramo, G., D'Angelo, C.A., Gzoyan, E. *et al.* Benchmarking research performance in a post-Soviet science system: the case of Armenia. *Scientometrics* 130, 2213–2235 (2025). https://doi.org/10.1007/s11192-025-05312-3

Balázs Schlemmer and Wolfgang Glänzel (2009). "Science in a Changing Europe: East Vs. West National Scientific Profiles by Subject Fields," *ISSI Newsletter* 5 (3): 52-58.

Chagelishvili, A., & Mushkudiani, Z. (2025). Georgia's Post-Independence Scientific Output and Prospects. *International Journal of Multidisciplinary: Applied Business and Education Research*, *6*(3), 1274-1291. https://doi.org/10.11594/ijmaber.06.03.21

Chankseliani, M., & Silova, I. (2018)." Reconfiguring Education purposes, policies, and practices during post-socialist transformations: Setting the stage." In M. Chankseliani & I. Silova (Eds.), *Comparing Post-Socialist Transformations: Purposes, Policies, and Practices in Education*, 7–25

Chankseliani, M., Lovakov, A. & Pislyakov (2021). V. A big picture: bibliometric study of academic publications from post-Soviet countries. *Scientometrics* 126, 8701–8730. https://doi.org/10.1007/s11192-021-04124-5

Gzoyan EG, LA Hovhannisyan, SA Aleksanyan, NA Ghazaryan, Sh.A Sargsyan (2015). "Comparative analysis of the scientific output of Armenia, Azerbaijan and Georgia," *Scientometrics* 102, 195-212.

Gzoyan, Edita, et al. (2023). "International visibility of Armenian domestic journals: the role of scientific diaspora." *Journal of Data and Information Science*, vol. 8, no. 2, 93-117. https://doi.org/10.2478/jdis-2023-0011

Humbatova Sugra Ingilab and Solmaz Aghazaki Abidi (2021). "The Current Position of Science Development in the World and in Azerbaijan," *Turkish Journal of Computer and Mathematics Education* 12 no. 6, 1356-1362

Mikail, E., Atun, Y. and Atun, A. (2019) Turkey-Azerbaijan Economical and Political Relations. *Open Journal of Political Science* 9, 512-524. doi: 10.4236/ojps.2019.93029

Moed H. F., de Moya-Anegon, F., Guerrero-Bote V., Lopez-Illescas C., & Hladchenko M. (2021). Bibliometric assessment of national scientific journals. Scientometrics, 126(4), 3641–3666. https://doi.org/10.1007/s11192-021-03883-5

Sargsyan Sh, DA Maisano, AR Mirzoyan, AA Manukyan, EG Gzoyan (2020). "EU-EAEU dilemma of Armenia: Does science support politics?" *Scientometrics* 122, 1491-1507.

Sukiasyan, N., & Davtyan, E. (2025). The South Caucasus Reconstructed: Polarity and Regional Security Order after the Nagorno-Karabakh War in 2020. *The International Spectator*, 1–18. https://doi.org/10.1080/03932729.2025.2500407

Yemelianova G. (2023). Turkey, the Karabakh Conflict and the Legacy of the Eastern Question. *Caucasus Survey*, 12(1), 73-102. https://doi.org/10.30965/23761202-bja10020

https://doi.org/10.51408/issi2025_056

Crossing Disciplinary Borders: How Italian SSH Journal Rankings Address Multidisciplinarity

Tindaro Cicero[1], Marco Malgarini[2], Marilena Maniaci[3]

[1]*tindaro.cicero@unimercatorum.it*
Universitas Mercatorum, Department of Engineering and Science, Piazza Mattei 10, 00186 Rome (Italy)

[2]*marco.malgarini@anvur.it*
ANVUR, Research Evaluation Department, Via Ippolito Nievo 35, 00100 Rome (Italy)

[3]*marilena.maniaci@anvur.it*
ANVUR, Governing Board, Via Ippolito Nievo 35, 00100 Rome (Italy)
University of Cassino and Southern Lazio, Department of Humanities, Campus Folcara, 03034 Cassino, FR (Italy)

Abstract

The classification of Italian academic journals in the Social Sciences and Humanities (SSH), carried out by ANVUR (the National Agency for the Evaluation of Universities and Research Institutes), plays a critical role in evaluating research output and shaping academic careers. However, the extent to which this classification adequately accounts for multidisciplinarity—a fundamental aspect of addressing complex societal challenges—remains underexplored. By analyzing the Italian classification framework and reviewing journal profiles, we identify systemic biases, disciplinary boundaries, and structural constraints that may hinder the integration of cross-disciplinary scholarship in the Italian academic landscape. Our findings reveal a partial and inconsistent consideration of this dimension, highlighting both recent advancements and persistent limitations in fostering cross-disciplinary collaboration among researchers. This study contributes to the ongoing debate on research evaluation in SSH, offering recommendations to improve classification systems so they better align with the evolving nature of scholarly inquiry, societal needs, and global research trends.

Introduction

Journal classification plays a crucial role in the evaluation of academic research, particularly in the social sciences and humanities (SSH) (Pontille & Torny, 2010; De Filippo et al., 2020; Cicero & Malgarini, 2020; Bonaccorsi et al, 2016; Ferrara and Bonaccorsi, 2016, Sivertsen, 2016). Nevertheless, traditional classification systems tend to prioritize monodisciplinary approaches, which can hinder the recognition of innovative research that spans multiple fields (Frodeman et al., 2017; Rafols et al., 2012). In Italy, the National Agency for the Evaluation of Universities and Research Institutes (ANVUR) is responsible for overseeing the classification of SSH journals. ANVUR's classification system is pivotal for assessing the quality of research outputs in the SSH sectors, serving as a benchmark for evaluating the quality of publications submitted by researchers for habilitation and as a key determinant in academic promotions. Indeed, to obtain the Italian National Scientific Qualification (Abilitazione Scientifica Nazionale, ASN), researchers must meet predefined thresholds based on the number of articles published in ANVUR-classified journals. ASN is structured around highly sectoralized disciplinary fields (Gruppi Scientifico-

Disciplinari, formerly Settori Concorsuali), making it challenging for scholars engaged in interdisciplinary research to gain appropriate recognition. In fact, academic career progression is strongly tied to fulfilling specific disciplinary requirements, which may disadvantage those whose work spans multiple fields. The Italian journal classification system provides a structured framework for researchers to identify reputable publishing venues and ensure transparency in research evaluation, but it has also been criticized for its rigidity, particularly regarding research that transcends traditional disciplinary boundaries.

First of all, it is essential to draw a clear distinction between multidisciplinarity and interdisciplinarity. Multidisciplinary refers to the concomitant use of multiple disciplines to address a scientific problem or to their coexistence within a single context, such as a journal, where each discipline maintains its distinct methodologies and perspectives. In this sense, a multidisciplinary journal can be identified by analyzing the range of disciplinary fields associated with it in the ANVUR classification system. For journals indexed in Scopus, this can be assessed through the subject categories assigned to each journal. In contrast, interdisciplinarity involves the integration of methods, theories, and frameworks from different disciplines to address complex problems (Klein, 1990). It transcends mere juxtaposition, fostering a synthesis that creates new knowledge or solutions However, assessing the degree of interdisciplinarity remains challenging—not only due to data limitations but also because it requires a deeper analysis beyond surface-level classifications, often involving complex peer review processes, the absence of established metrics, and the constraints of rigid disciplinary boundaries While multidisciplinarity and interdisciplinarity are distinct concepts, they are nonetheless interrelated, as the integration of multiple disciplines often serves as a foundation for deeper interactions and synthesis across fields.

While relevant distinctions have been made in the literature, a systematic investigation of multidisciplinarity within the Social Sciences and Humanities (SSH), particularly in the Italian context, remains largely unexplored. Several studies have examined the extent and modalities through which fields in both SSH and the Science, Technology, and Medical (STM) domains engage in multidisciplinary practices. Notably, Soós et al. (2018) contest the conventional dichotomy between SSH and STM, showing that certain fields across these domains exhibit significant overlaps in their multidisciplinary profiles, thereby challenging the "two cultures" thesis. Their findings suggest that multidisciplinarity varies not only across disciplines but also along different analytical dimensions, pointing to the need for a more nuanced conceptualization. Moreover, the study argues that SSH and STM should be understood as umbrella categories—useful for administrative and communicative purposes, yet misaligned with the actual cognitive and structural organization of science. In parallel, other contributions have explored the institutional and epistemic challenges in integrating SSH into interdisciplinary research funding frameworks (Pedersen, 2016; Välikangas, 2024; Gerli, 2020). Additionally, some studies have assessed the degree of multidisciplinarity at the journal level across broad comparative datasets (Redondo-Gómez et al., 2024), though often without a dedicated focus on SSH, or concentrating on highly

multidisciplinary journals such as Nature or Science (Ackerson & Chapman, 2023; Solomon et al., 2016; Ding et al., 2018).

This paper presents a preliminary analysis aimed at uncovering structural limitations in the way multidisciplinarity is recognized—or neglected—within existing journal classification systems. Specifically, it seeks to investigate whether, and in what ways, multidisciplinary research is adequately acknowledged and represented in such frameworks.

It seeks to explore whether and how multidisciplinary research is acknowledged within ANVUR's journal classification system. Specifically, it investigates the extent to which journals that engage with multiple disciplines in the SSH are classified and valued, and how this impacts the visibility and evaluation of multidisciplinary scholarship. Addressing this gap, the paper contributes to the broader debate on research evaluation in SSH, providing empirical insights into whether current classification practices facilitate or constrain interdisciplinary scholarship in Italy. Although this study focuses on journal-level classifications, future research could extend the analysis of interdisciplinarity to the article level. One promising approach is to examine co-authorship networks, identifying collaborations between researchers from different disciplines as a proxy for interdisciplinary engagement. By mapping these networks, it may be possible to identify patterns of knowledge exchange, disciplinary integration, and the emergence of cross-field collaborations, especially considering how the structure and dynamics of research collaborations have evolved in recent years. Nevertheless, this approach should be used with caution, as it may have limitations: citation-based tools are often inadequate in the SSH due to lower citation rates and the prevalence of monographs or publications in national languages. This shift has been driven by the increasing recognition that complex global challenges – such as climate change, health crises, artificial intelligence, and social inequalities – require cross-disciplinary solutions. This evolution is particularly evident in competitive funding programs at both national and international levels. Additionally, analyzing citation networks could provide insights into how interdisciplinary work is received and integrated within academic discourse, shedding light on the real impact of cross-disciplinary research in SSH.

The structure of this paper is as follows: Section 2 provides an overview of ANVUR's journal classification system and its criteria. Section 3 examines the representation of multi- and interdisciplinary journals within the classification. Section 4 discusses the implications of these findings for the recognition of interdisciplinary research in SSH, offering a preliminary set of conclusions and recommendations for improving the evaluation of interdisciplinary scholarship.

The ANVUR classification system

Since 2012, the National Agency for the Evaluation of Universities and Research Institutes (ANVUR) has been tasked to maintain a list of scientific and top tier ('A-Class') journals, to be used by the Ministry of University in the context of the National Scientific Qualification procedures for social sciences and humanities. The classification is ruled by a specific regulation, delineating the criteria, parameters,

and procedures for classifying and updating the lists. Evaluation is performed by a specifically designated committees of professors, whose members are selected by drawing lots from a list defined based on a public call for expressions of interest. Experts included in the list should possess high scientific qualifications and adequate experience in evaluation. The classification of journals is used both as a tool to determine the qualification of prospective committee members and to define the eligibility thresholds for candidates. The regulation specifies detailed criteria for classifying journals, including considerations of scientific relevance, originality, peer-review processes, editorial quality, and adherence to ethical standards. The classification aims to ensure that journals meet rigorous academic standards and contribute meaningfully to their respective fields.

As per 2024, ANVUR has classified a total of over 23,000 journals. Table 1 shows their distribution across 6 different disciplinary areas (Architecture; Classical Studies, Philology and Literatures, History of Art; History, Philosophy and Education; Law; Economics and Statistics; Social and Political Sciences), also providing information about the share of journals classified by ANVUR which are also indexed in Scopus. Data shows that, overall, over 54% of journals included in the ANVUR lists of scientific journals is also indexed in Scopus, the share rising to 73% for top-tier (A-class) journals. Indexation is particularly common in Economics and Statistics, Social and Political Sciences and History, Philosophy and Education, being on the other hand less relevant in Law and in the diverse field of 'literary studies'.

Table 1. Classified journals by disciplinary field.

Area	N. of scientific journals	Of which: indexed in Scopus	N. of A-Class journals	Of which: indexed in Scopus
Architecture	2,547	1,186	451	336
Literary studies	7,803	3,379	2,758	1,718
History, Philosophy, Education	8,636	4,698	2,270	1,818
Law	2,851	931	734	309
Economics and statistics	8,265	5,883	1,460	1,415
Social and political sciences	5,414	3,222	1,755	1,491
Total	23,456	12,620	7,775	5,651

Multidisciplinarity of the ANVUR classification

Multidisciplinarity involves the collaboration of multiple disciplines to address a shared problem or topic, where each field retains its methods and perspectives without integrating them. Multidisciplinary approaches are often employed in addressing complex challenges like public health, urban planning, or climate change, benefiting from the diverse expertise of various fields (Max-Neef, 2005). While

multidisciplinarity fosters creativity and efficiency, it can result in fragmented insights and communication barriers (Tress et al., 2005). For example, multidisciplinary research on sustainable development might involve economists, ecologists, and sociologists working independently to contribute to holistic solutions. Though limited in integration, this approach may still be functional, and even crucial, in addressing broad, multifaceted global issues.

In the following analysis, we will examine multidisciplinarity through two complementary approaches. The first approach is based on the Scopus classification, thus considering only the share of journals indexed in this database, leveraging the ASJC (All Science Journal Classification) system.

The choice to focus on Scopus-indexed journals is driven by both methodological and regulatory considerations. First, inclusion in Scopus constitutes a sufficient condition for a journal to be recognized as scientific under current ANVUR regulations. Second, the ASJC (All Science Journal Classification) system provides a well-established and structured framework for identifying multidisciplinarity at the journal level, which aligns with the unit of analysis adopted by the ANVUR classification system. At this stage, alternative databases—such as the open-access platform OpenAlex—have not been considered, as they do not offer a classification of journals by disciplinary area but instead assign topics at the article level, making them less suitable for the purposes of this study.

The second approach relies on the ANVUR classification, examining the simultaneous presence of journals across multiple areas. This dual perspective will allow us to capture different dimensions of multidisciplinarity and assess its relevance in academic publishing.

Multidisciplinarity in indexed journals: an analysis based on the Scopus classification (ASJC)

In order to provide some first evidence about the degree of multidisciplinarity of the journals classified by ANVUR, for each scientific and A-Class journal included in our list that is also indexed in Scopus the number of ASJC in which it is included has been calculated. (Table 2): the higher the share of journals that are indexed in a high number of ASJC, the more the ANVUR classification in that particular field may be considered as multidisciplinary, in the sense defined above. The ANVUR database is updated as in the Spring of 2024, while the Scopus title list used in the analysis is that of December 2024.

Multidisciplinariy is particularly important in A-Class journals belonging to the broad field of ‘literary studies’: over 65% of the top-tier journals in this field are indeed indexed in at least two Scopus categories. Multidisciplinarity is also widespread among A-Class journals in Economics and Statistics and among both scientific and A-Class journals in Architecture. On the other hand, Law journals are mostly monodisciplinary, in the sense that 60% of those that are indexed in Scopus are classified only in one ASJC, and 93% of them are at most classified in two ASJC. Interestingly, in Architecture, History, Philosophy and Education and Social and Political Sciences the share of multidisciplinary journals is similar for scientific and A-Class journals. On the other hand, in Economics and Statistics and, to a minor

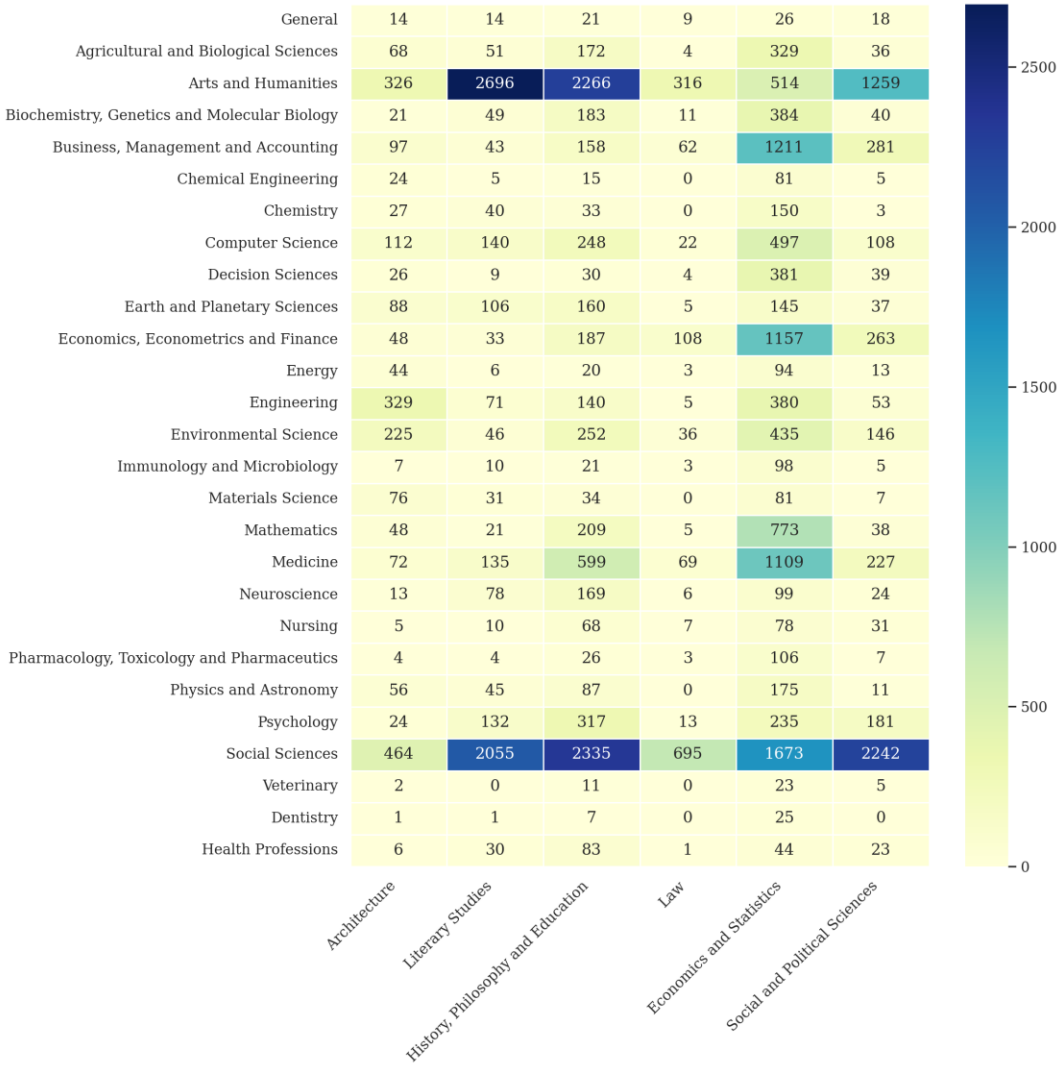
degree, in 'literary studies' multidisciplinarity is especially found in A-Class journals rather than in those recognised solely as scientific.

Table 2. Multidisciplinarity of ANVUR' journals with respect to the ASJC classification.

	Architecture		Literary studies		History, Philosophy and education		Law		Economics and statistics		Social and Political Sciences	
Number of AJC	Scientific	A Class	Scientific	A Class	Scientific	A Class	Scientific	A Class	Scientific	A Class	Scientific	A Class
1	38,4%	37,5 %	37,7%	33,3 %	46,5%	45,6%	59,8%	63,1%	45,5%	35,7%	52,8%	54,7%
2	40,6%	44,0 %	54,1%	59,9 %	42,9%	44,2%	33,1%	32,7%	38,7%	45,1%	38,7%	37,6%
3	16,5%	13,7 %	6,1%	5,5%	8,3%	7,7%	5,7%	3,6%	12,0%	13,6%	6,3%	5,9%
4	3,5%	4,2 %	1,4%	0,8%	1,8%	2,1%	1,1%	0,6%	2,8%	5,0%	1,8%	1,4%
5	0,8%	0,6 %	0,4%	0,3%	0,4%	0,4%	0,3%	0,0%	0,6%	0,6%	0,3%	0,5%
6	0,1%	0,0 %	0,2%	0,1%	0,1%	0,0%	0,0%	0,0%	0,2%	0,0%	0,1%	0,0%
7	0,0%	0,0 %	0,1%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
8	0,0%	0,0 %	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%

In which subject categories are multidisciplinary journals mostly indexed? Figure 1 provides an answer to this question, showing, for each scientific area, the "map" of where journals are most concentrated in terms of indexation categories. Journals classified in the ANVUR lists and indexed in Scopus appear particularly in the ASJC categories "Social sciences" and "Arts and humanities", as expected. Journals in Economics and Statistics show however a more pronounced multidisciplinary profile, with a considerable number of journals pertaining to different Scopus domains. In History, Philosophy and Education, scientific journals include a relevant component indexed in the medical area, mostly associated with neurosciences and psychology, epidemiology and public health. In Architecture, 'Literary Studies' and Law most journals are indexed in the expected categories ("Arts and Humanities" and "Social Sciences").

a) Scientific Journals



b) Class A Journals

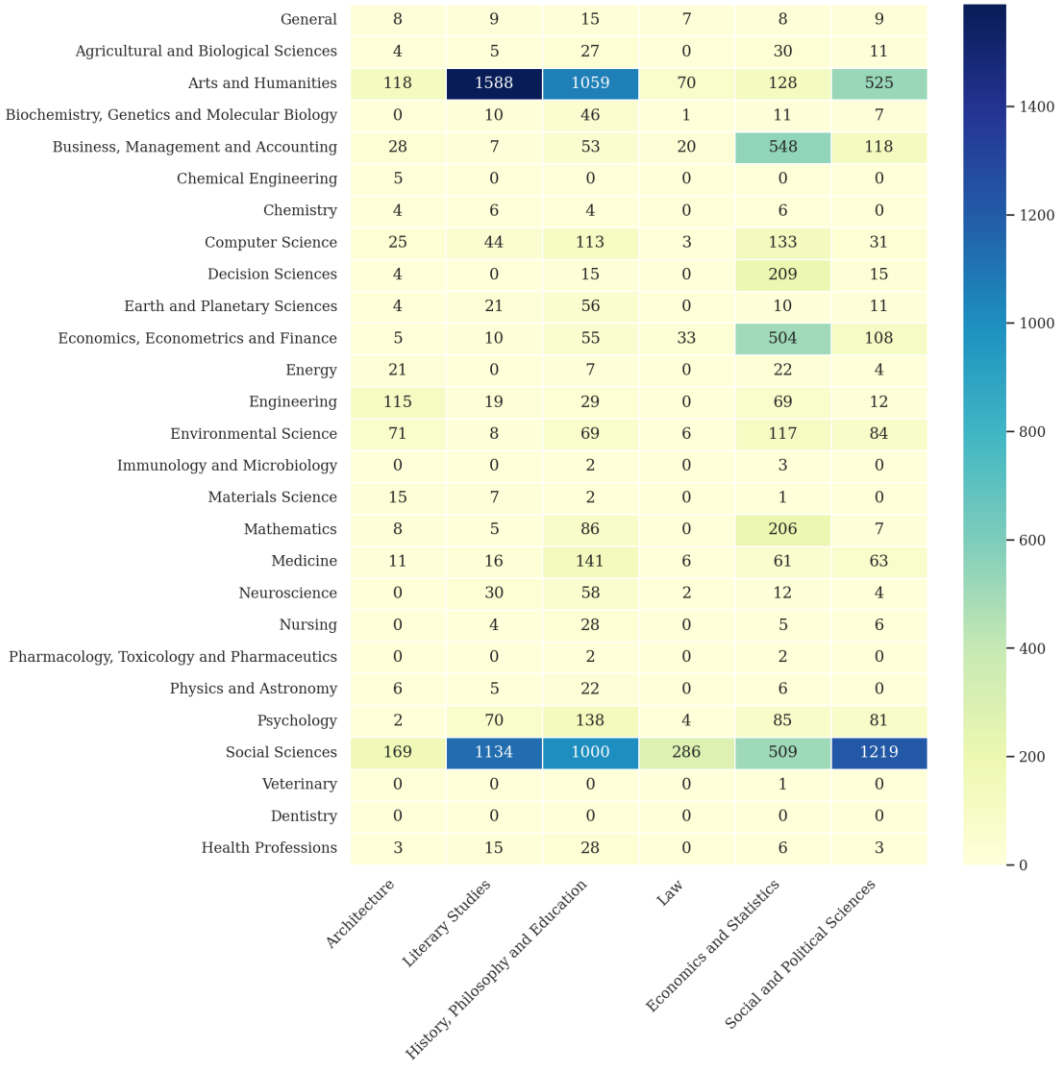


Figure 1. Journals by disciplinary field and ASJC code.

Multidisciplinarity in ANVUR's classification: assessing the presence across multiple areas

A similar analysis may be performed to check whether ANVUR journals are classified in only one or multiple Italian research areas. Table 3 shows that over 65% of scientific journal are indeed monodisciplinary, i.e., they are classified only in one of the 6 disciplinary areas of interest; on the other hand, a very limited number of journals is fully multidisciplinary, i.e., it is classified in all the areas (0,2%). However, overall, 35% of scientific journals are indeed classified at least in two areas, showing a remarkable degree of multidisciplinarity of the classification. On the other hand, the A-Class classification is more discipline-specific: only 17% of journals are indeed recognised as A-Class in more than one area, probably also since

the classification serves the scope of identifying adequate candidates for the National Scientific Qualification procedure, which is indeed granted on a disciplinary basis.

Table 3. Multidisciplinary of ANVUR' journals with respect to the Italian disciplinary fields.

Classification of journals by disciplinary areas	N. of scientific journals	N. of A-Class journals
Journals classified in one disciplinary area	15,373	6,396
Journals classified in two disciplinary areas	5,271	1,161
Journals classified in three disciplinary areas	1,927	176
Journals classified in four disciplinary areas	657	35
Journals classified in five disciplinary areas	176	0
Journals classified in six disciplinary areas	52	7
Total	23,456	7,775

Finally, tables 4 and 5 provide information about overlapping classification among disciplines for scientific and A-class journals, respectively: each cell of the tables report the share of journals that are classified in both the row and column disciplines, the share of journals that are classified only in one discipline being represented on the diagonal of the matrix. The colour scale (blue for scientific journals, red for A-Class) visually highlights the cases of major interchange among disciplines. Most scientific journals are monodisciplinary in all areas (table 4); however, journals classified in Architecture are often classified also in other areas (but not in Law); 'Literary studies' journals are also found in History, Philosophy and Education, and vice versa, with the latter discipline being also interrelated with Social and Political Sciences. Law journals are indeed classified also in other disciplines but in Architecture. Most journals in Economics and Statistics are monodisciplinary, with some interchange with History, Philosophy and Education and Social and Political sciences. Lastly, Social and Political Sciences journals are more multidisciplinary with respect to those of other areas, being most of the time classified also in other disciplines.

A-Class journals are more disciplinary concentrated, as results from by the high proportion of journals on the main diagonal of the matrix (table 5). More specifically, around ¾ of A-Class journals are indeed monodisciplinary in Literary Studies and Law, and over 60% in Economics and Statistics. Most journals are instead at least present in two areas in Architecture and Social and Political Sciences. Interchange is particularly strong among History, Philosophy and Education and Social and Political Sciences, and among the latter and Economics and Statistics. Some mutual recognition of A-Class journals also occurs in Literary Studies and History, Philosophy and Education. Architecture show moderate commonalities in journals

classification with all the other areas, with no particular area emerging. Finally, overlap with other disciplines seldom occurs in Law, with however just about 10% of journals being also classified in Social and Political Sciences.

Table 4. Overlap matrix of scientific journals.

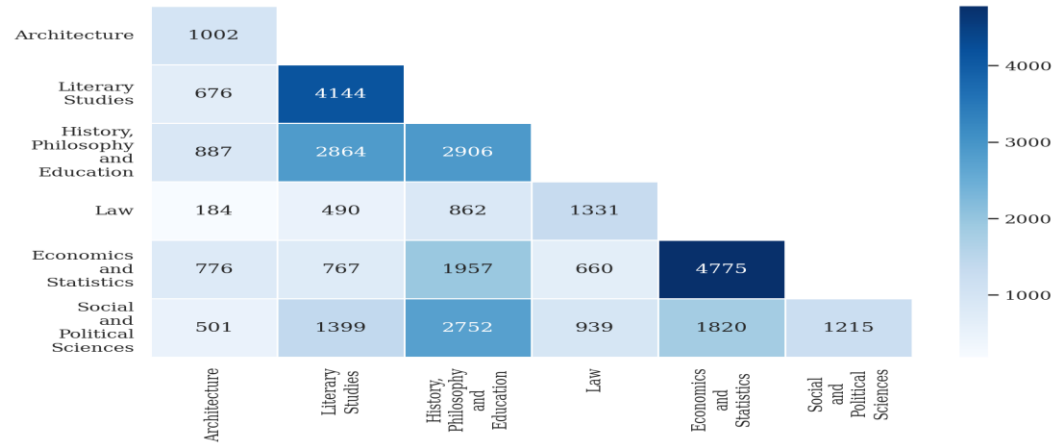
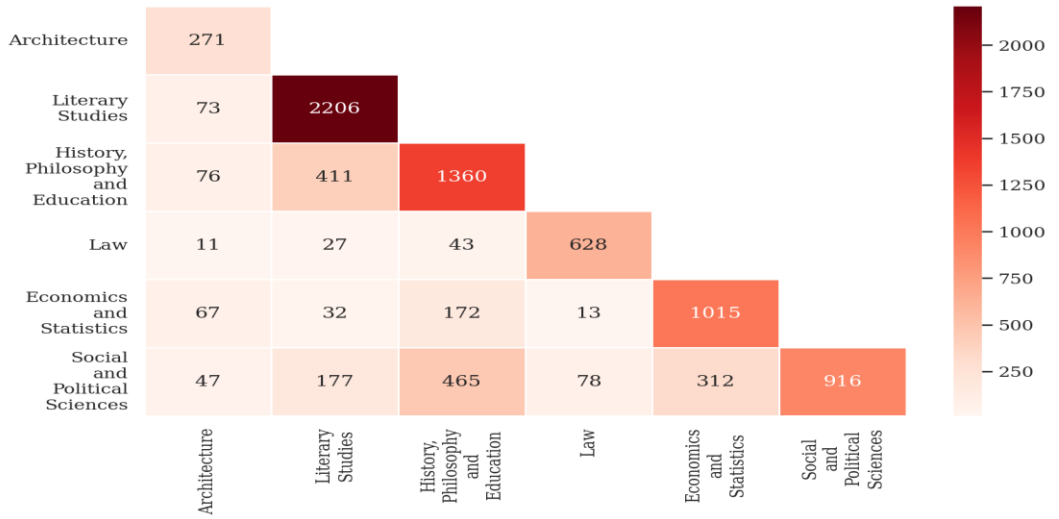


Table 5. Overlap matrix of A-Class journal.



Limitations and Future Research

This study presents a preliminary and journal-level analysis of how multidisciplinarity is acknowledged within the Italian SSH journal classification system. While it provides meaningful insights into the structural and disciplinary dynamics of the ANVUR framework, several limitations must be acknowledged. First, the reliance on Scopus and its ASJC classification system, while methodologically justified, excludes journals not indexed in this database, potentially overlooking relevant outlets, particularly those published in national

languages or with limited international circulation. Moreover, the analysis focuses exclusively on journal-level metadata, without assessing the content or citation patterns of individual articles, which might offer a more nuanced understanding of interdisciplinary practices.

It should also be noted that, on the basis of the available data, the degree of interdisciplinarity cannot be appropriately addressed, as it requires a more refined analysis at the level of individual research outputs and within specific collaborative research networks—an avenue we leave for future research. To assess interdisciplinarity at the level of individual publications, it might be fruitful to conduct citation-based analyses, investigating whether an article references sources from diverse disciplines or is cited by scholars from different fields, suggesting cross-disciplinary impact. Similarly, semantic analysis using text-mining techniques could help reveal whether a publication integrates theories, methodologies, or conceptual frameworks from multiple disciplines, thereby offering further validation of its interdisciplinary nature.

Beyond the level of individual publications, collaborative networks offer another valuable lens through which interdisciplinarity can be assessed. Co-authorship network analysis—applied to datasets such as the national academic and research information system managed by the Italian Ministry of University and Research (LoginMIUR)—could be used to identify patterns of collaboration among researchers from different disciplinary backgrounds. For example, frequent co-authorship between scholars affiliated with distinct research areas may reflect an interdisciplinary research environment. Analyzing institutional affiliations may also shed light on whether such collaborations occur across departments or institutions, thus reflecting broader structural trends. Network metrics such as degree centrality (measuring the extent of a researcher's collaborative connections) and betweenness centrality (indicating the extent to which a researcher bridges different communities) could help quantify the role of interdisciplinary collaborations in shaping the scientific landscape. Such metrics may highlight whether cross-disciplinary interactions are occurring between traditionally separated domains—e.g., between the humanities and hard sciences—or within subfields of a single domain.

Future research adopting these complementary approaches would allow for a more comprehensive evaluation of interdisciplinarity, moving beyond the limitations of rigid journal-based classifications and towards a more dynamic and integrative understanding of how interdisciplinary knowledge is produced and disseminated within the Italian SSH landscape.

Conclusions

The analysis conducted in this paper reveals the structural tensions and limitations embedded in the current Italian journal classification system when it comes to the recognition and valorization of multidisciplinary research in the Social Sciences and Humanities (SSH). While ANVUR's classification provides a crucial framework for academic evaluation - particularly through its influence on habilitation procedures and career advancement - it remains strongly anchored to a disciplinary logic that does not fully accommodate the evolving nature of contemporary scholarly inquiry.

The study shows that multidisciplinarity is only partially reflected in the classification system. On one hand, there is clear evidence that a non-negligible share of journals, especially among those indexed in Scopus and particularly in fields such as Economics and Literary Studies—do engage with multiple subject categories, suggesting an openness to cross-disciplinary perspectives. On the other hand, this recognition is uneven and often limited to scientific journals rather than A-Class ones, which are more strictly bound to disciplinary criteria, likely reflecting the sectorialized structure of the National Scientific Qualification process. Fields such as Law and, to a lesser extent, Literary Studies show a marked tendency toward monodisciplinarity, potentially narrowing the space for cross-boundary dialogue and innovation.

Moreover, the use of ASJC codes from Scopus as a proxy for multidisciplinarity, while methodologically sound and aligned with regulatory constraints, also reveals the dependency of national evaluation systems on bibliometric infrastructures that may not be fully suited to the specificities of SSH research. In this context, the lack of a dedicated mechanism for capturing multidisciplinarity at the journal level within the ANVUR framework emerges as a critical gap. The tendency to evaluate research outputs through the lens of disciplinary classifications may inadvertently penalize those contributions that, by their very nature, do not fit neatly within established academic boundaries.

From a policy perspective, the findings of this study call for a reconsideration of the principles and procedures underpinning journal classification in SSH. In particular, there is a need to introduce greater flexibility in how multidisciplinarity is assessed and rewarded. This may include recognizing journals with broader disciplinary scope, facilitating multiple area classifications more systematically, and reducing the weight of strict disciplinary silos in research evaluation. Failure to address these issues risks reinforcing an academic ecosystem that is ill-equipped to respond to the complex societal challenges that demand integrative, cross-disciplinary approaches.

Ultimately, while the current classification system provides a necessary structure for the governance of academic evaluation, it should evolve to reflect the increasingly hybrid and dynamic nature of SSH scholarship. A more inclusive and nuanced approach to multidisciplinarity would not only enhance the fairness and accuracy of evaluation procedures but also contribute to fostering a research environment that supports innovation, dialogue across disciplines, and the production of knowledge that is both academically robust and socially relevant.

References

Ackerson, L. G., & Chapman, K. (2003). Identifying the role of multidisciplinary journals in scientific research. College & Research Libraries, 64(6), 468-478.

Bonaccorsi A, Cicero T, Ferrara A and Malgarini M. Journal ratings as predictors of articles quality in Arts, Humanities and Social Sciences: an analysis based on the Italian Research Evaluation Exercise *F1000Research* 2015, 4:196, https://doi.org/10.12688/f1000research.6478.1)

Bonaccorsi, A. (2018). Addressing the disenchantment: Universities and regional development in peripheral regions. *Regional Studies*, 52(8), 1025-1035.

Cicero, T., Malgarini, M. (2020). On the use of journal classification in social sciences and humanities: evidence from an Italian database. Scientometrics 125, 1689–1708, https://doi.org/10.1007/s11192-020-03581-8

De Filippo, D., Aleixandre-Benavent, R., & Sanz-Casado, E. (2020). Toward a classification of Spanish scholarly journals in social sciences and humanities considering their impact and visibility. *Scientometrics*, 125(2), 1709-1732.

Ding, J., Ahlgren, P., Yang, L., & Yue, T. (2018). Disciplinary structures in Nature, Science and PNAS: Journal and country levels. Scientometrics, 116, 1817-1852.

Ferrara, A., Bonaccorsi, A. (2016). How robust is journal rating in Humanities and Social Sciences? Evidence from a large-scale, multi-method exercise. Research Evaluation, 25(3), 279-291.

Frodeman, R., Klein, J. T., & Mitcham, C. (Eds.). (2017). *The Oxford Handbook of Interdisciplinarity*. Oxford University Press.

Gerli, M. (2020). Where are the social sciences going to? The case of the EU-Funded SSH Research Projects. In Text Analytics: Advances and Challenges (pp. 225-240). Springer International Publishing.

Klein, J. T. (1990). *Interdisciplinarity: History, Theory, and Practice*. Wayne State University Press.

Max-Neef, M. A. (2005). Foundations of transdisciplinarity. Ecological Economics, 53(1), 5-16.

Pedersen, D. B. (2016). Integrating social sciences and humanities in interdisciplinary research. *Palgrave Communications*, *2*(1), 1-7.

Pontille, D., Torny, D. (2010). The controversial policies of journal ratings: evaluating social sciences and humanities. *Research Evaluation*, 2010, 19 (5), 347-360.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. *Research Policy*, 41(7), 1262-1282.

Redondo-Gómez, D., Arroyo-Machado, W., Torres-Salinas, D., Margalida, A., & Moleón, M. (2024). A long-term assessment of the multidisciplinary degree of multidisciplinary journals. PloS one, 19(12), e0314616.

Sivertsen, G. (2016). Patterns of internationalization and criteria for research assessment in the social sciences and humanities. *Scientometrics*, 107, 357-368.

Solomon, G. E., Carley, S., & Porter, A. L. (2016). How multidisciplinary are the multidisciplinary journals science and nature?. *PloS one*, *11*(4), e0152637.

Soós, S., Vida, Z., & Schubert, A. (2018). Long-term trends in the multidisciplinarity of some typical natural and social sciences, and its implications on the SSH versus STM distinction. *Scientometrics*, *114*, 795-822.

Tress, B., Tress, G., & Fry, G. (2005). Defining concepts and the process of knowledge production in integrative research. *Environmental Management*, 36(1), 1-14.

Välikangas, A. (2024). The Limited Role of Social Sciences and Humanities in Interdisciplinary Funding: What are Its Effects?. Social Epistemology, 38(2), 152-172.

https://doi.org/10.51408/issi2025_015

Digital Twins in Healthcare: State of the Art, Bibliometric Analysis and Future Perspectives

Cinzia Daraio[1], Simone Di Leo[2], Jacopo Orsini[3]

[1] *daraio@diag.uniroma1.it*, [2]*dileo@diag.uniroma1.it*, [3]*orsini.2099929@studenti.uniroma1.it*
DIAG Sapienza University of Rome, via Ariosto, 25, I-00185 Rome (Italy)

Abstract

Digital Twins (DTs) are reshaping healthcare by providing dynamic, digital counterparts of physical systems, enabling real-time interaction, simulation, and analysis. These systems leverage advanced modeling and real-world data integration to optimize medical training, planning, and patient-specific care. This paper explores the evolution of DTs in healthcare, presenting a bibliometric analysis of trends and outlying future directions.

The state-of-the-art section highlights technological advances in DTs, with a particular focus on simulating complex physiological behaviors. These advancements align with the growing demand for precision in surgical training and planning. The bibliometric analysis reveals an exponential increase in research interest, driven by advancements in Artificial Intelligence (AI), immersive technologies, and real-time data processing. Cross-disciplinary efforts, combining fields such as computer science, biomechanics, and medical engineering, are highlighted as key enablers, expanding the applicability of DTs. Future perspectives emphasize the transformative potential of DTs in remote surgical procedures, augmented diagnostics, and personalized medicine. The integration of Augmented Reality (AR) and Virtual Reality (VR) enhances the user experience by providing immersive, interactive environments. Additionally, the inclusion of haptic feedback and sensor-based tracking further augments realism, improving usability and adoption. Through the case study of the Rome Technopole project "*Phygital Twin Technologies for Innovative Surgical Training & Planning*", this paper showcases how DTs are already impacting healthcare, from training simulators to patient-specific planning tools. Furthermore, the discussion points to new frontiers, such as integrating predictive analytics for proactive healthcare interventions.

Introduction and aim

The healthcare field has been profoundly transformed by technological advancements over the past few decades. These innovations have paved the way for numerous new frontiers, each reshaping how medical professionals diagnose, treat, and manage various health conditions.

Digital Twins (DTs) are at the forefront of transformative innovations in healthcare, offering groundbreaking applications in training, planning, and personalized medicine. These advanced digital counterparts of physical systems allow real-time interaction and data-driven decision-making.

Phygital Twin technologies represent an advanced evolution of next-generation enabling technologies, building upon the foundational concept of DTs. DTs have already demonstrated their versatility across a wide range of applications, including Industry 4.0 and Connected Health (Pires et al., 2019; Bagaria et al., 2020; Evangeline, 2020; Aziz et al., 2024; El-Agamy et al., 2024). The Phygital approach emphasizes that DTs should not only replicate, monitor, predict, and optimize the processes and characteristics of their physical counterparts, referred to as Physical Twins, but also maintain real-time interconnectivity (Grieves & Vickers, 2017; Jones

et al., 2020; Mourtzis et al., 2023; van Dinter et al., 2022). The rapid growth of big data and continuous advancement in data science and artificial intelligence have the potential to significantly advance DTs and phygital research and development. Although various DTs initiatives have been underway in the industrial sector, DTs for health are still in their early stages **(**Katsoulakis et al., 2024**).**

This paper provides an in-depth exploration of the current landscape, a review of bibliometric trends, and identifies venues for future directions. Furthermore, by applying established bibliometric techniques to the emerging and interdisciplinary field of DTs in healthcare, this study offers insights into the challenges and specific characteristics of analyzing the scientific literature in such rapidly evolving domains. It aligns with the goals of the Rome Technopole project "Phygital Twin Technologies for Innovative Surgical Training & Planning".

The state-of-the-art technologies discussed in the paper highlight a critical shift from traditional simulation methodologies like Finite Element Methods (FEM) to newer paradigms such as Physics-Informed Neural Networks (PINNs). FEM, while reliable, struggles with the computational demands of real-time simulation, which is essential for applications involving interactive digital twins. PINNs, on the other hand, offer a promising alternative by enabling rapid, accurate modeling of complex systems, including deformable tissues. This innovation is particularly relevant to the Rome Technopole project's work on developing a Digital Phantom (DP), a digital replica of anatomical structures designed to enhance surgical training. By simulating realistic tissue behavior in response to external forces, the DP promises to transform how surgeons learn and practice their craft (see Distefano et al., 2023; De Santis et al., 2024).

An integral part of this advancement is the immersive environment created through a seamless combination of technologies. Virtual Reality (VR) and augmented reality (AR) frameworks are employed to offer practitioners a fully interactive experience. Using wearable haptic devices like WeART TouchDiver and precision tracking tools such as the NDI Polaris Vega XT, users can engage with digital models as if they were physical objects. This approach not only enhances the realism of training exercises but also provides additional sensory feedback, such as tactile sensations, visual cues, and real-time alarms, enriching the learning process.

The Rome Technopole project has gone a step further by integrating these capabilities into demonstrators showcasing the possibilities of DTs. These demonstrators use advanced physics engines like MuJoCo to simulate 3D deformations in anatomical models. By optimizing computational processes through multithreading and synchronization techniques, these simulations achieve the real-time responsiveness required for immersive VR applications. This marks a significant leap in usability and interaction quality, addressing one of the major limitations of earlier approaches. The main objective of this research is to map the technological frontiers related to DTs in the medical field, identifying the gaps in the literature, the main topics and clinical areas covered, and the trends and advantages and disadvantages of applying these technologies in the medical field. To address this topic, we provide a systematic literature review, supporting it with bibliometric analyses to assess trends and main topics of the state of the art literature. The paper

is organized as follows. The next section presents the systematic review approach. The following section reports the bibliometric analyses carried out. The next section illustrates the results of the systematic review, and the final section concludes the paper.

Methods

A systematic literature review (SLR) is a research method designed to precisely identify, evaluate, and summarize all relevant evidence on a specific research question on the base of systematic procedures to minimize human error and bias. The SLR carried out to identify how DTs have been used in the medical field, was performed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Page et al., 2021) framework. The applied procedure follows Avenali et al. (2023). Key steps in conducting a SLR include:

1. Establishing clear objectives and pre-defined eligibility criteria for study inclusion.
2. Employing an explicit and reproducible methodology.
3. Conducting a systematic search to identify all studies meeting the eligibility criteria.
4. Assessing and screening the identified studies.
5. Systematically presenting, describing, and summarizing the included studies.

These steps are designed to minimize bias and ensure robust and reliable results.

As eligibility criteria for the inclusion of the studies, this research adopts language, type of article and type of source. A keyword-based search was done on "Title", "abstract" and "keywords" using Scopus (https://www.scopus.com/, last accessed 17/01/2025) and Web of Science (https://www.webofscience.com/wos/woscc/basic-search, last accessed 17/01/2025) databases. We selected only reviews, chapter book and articles in English or in Italian, published in indexed journals or indexed books. The query executed for both databases is as follows:

"Digital twin" OR "Digital twins" OR "Digital phantom" OR "Digital phantoms" (All Fields)
AND **Practice OR Training OR "Teaching purpose" OR "Surgical training" OR "Robot assisted surgery"** (All Fields)
AND **healthcare OR "health care"** (All Fields)
AND NOT **"industry 5.0" OR "industry 4.0"** (All Fields)
AND **Article** or **Review Article** (Document Types)
AND **English** or **Italian** (Languages)

According to the reproducibility characteristic of SLR, all steps related to the skimming of articles are reported in detail. Initially, we obtained 54 articles from Scopus and 101 articles from Web of Science. After collection and subsequent selection of articles according to the search objective and predefined selection criteria we obtained 104 articles. After reading and deepening these articles, 32 articles were discarded because the main topic treated in the article was not in the healthcare field.

At the end of the screening and selection process, we obtained 72 articles, which will be analysed below. Figure 1 reports the PRISMA diagram, detailing the performed screening and selection procedure.
For each article retained, information was extracted on the technologies used or proposed, the advantages and disadvantages of these technologies and the clinical field in which they were used.

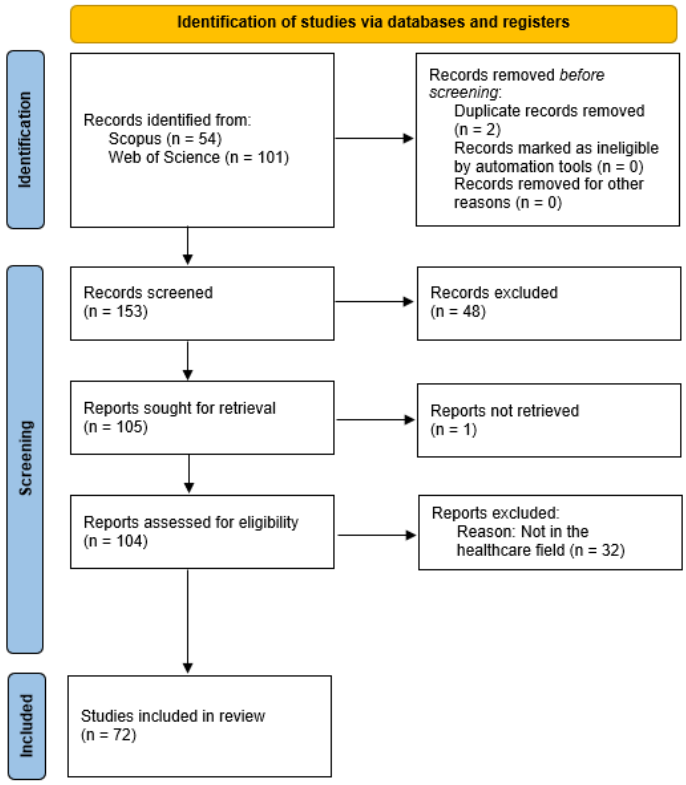


Figure 1. PRISMA 2020 flow diagram (Page et al. 2021).

To complement our SLR and gain a comprehensive understanding of the existing literature, we employed several bibliometric analyses. Combining bibliometric analysis with a SLR facilitates the efficient and reproducible generation of new knowledge from existing research (Avenali et al., 2023). Bibliometric analysis leverage articles metadata to uncover insights into various aspects, such as author collaborations, relationships between countries, and prominent authors (see e.g. Broadus, 1987). Specifically, our bibliometric analyses includes:

- An analysis of the relationships between authors' countries, keywords (representing key topics), and publication sources.
- An analysis of the frequency of author keywords.
- A factorial analysis using dimensionality reduction techniques on bigrams (sequences of two adjacent words) extracted from abstracts.
- A thematic analysis using bigrams from the abstract.

The relationships between authors' countries, keywords (representing key topics), and publication sources were analyzed using a Sankey diagram (Yang, 2022), a data visualization technique that illustrates associations between different article characteristics. This diagram visually represents the magnitude of flows between interconnected elements. The thickness of the links between nodes (representing keywords, countries, or publication sources in this study) is proportional to the volume of interactions between them. By mapping these flows, the analysis aimed to uncover patterns of international research collaboration, thematic clusters, and the influence of various publication sources on digital twin research in the healthcare field.

To further examine thematic trends, keyword occurrence analysis was conducted to determine the frequency and distribution of the most used terms in literature. Calculating keyword frequencies helped identify predominant research areas (Donthu et al., 2021). Beyond keyword analysis, bigram analysis of abstracts was performed to gain deeper insights into thematic connections and domain-specific vocabulary. Since keywords serve as proxies for the main topics of articles, analyzing abstract bigrams provided a more comprehensive understanding of underlying research themes. To explore key thematic areas, we applied Multiple Correspondence Analysis (MCA), a factorial and dimensionality reduction technique, to the abstract bigrams. MCA projects keywords into a two-dimensional space, revealing thematic clusters and underlying relationships within the field (Greenacre, 2017). This analysis highlighted distinct research themes, each representing a specific area of interest. Finally, co-word network analysis and clustering, following the method proposed by Cobo et al. (2011), were used for a thematic analysis of abstract bigrams. The identified clusters were positioned within a four-quadrant diagram based on their centrality and density, classifying them into central, niche, emerging/declining, or cross-cutting/basic themes. This approach helped delineate the primary research areas and the prevailing conceptual links within the field. All analyses were conducted using the R package Bibliometrix (Aria & Cuccurullo, 2017).

Results from the bibliometric analyses

This section presents the results of the bibliometric analyses conducted on the selected articles. Before delving into these findings, we provide a brief overview of the selected publications. A total of 72 articles, published between 2018 and 2024, were sourced from 53 journals and books. The field has experienced rapid growth, with an annual publication increase of 75.28%, highlighting its expanding significance. The average document age of 1.03 years reflects the field's recent rise in research interest, largely driven by the increasing focus on DTs.

Analysis of the relationships between authors' countries, keywords and sources

Figure 2 visually depicts the intricate relationships between countries, author keywords, and publication sources within the medical research landscape. The central column, dominated by keywords such as artificial intelligence, digital twins, and machine learning, highlights a strong focus on emerging technologies and their

applications in medicine. The results further illustrate the dynamic nature of research collaborations, as shown by the connections between various countries, keywords, and sources. Notably, the United States emerges as a key player, exhibiting extensive connections across diverse research areas. Likewise, Germany, the United Kingdom, and Switzerland demonstrate significant research activity and strong international collaborations. On the right-hand column, the sources primarily consist of scientific journals, reflecting the preferred publication venues for this research. The presence of journals such as IEEE Transactions on Consumer Electronics, Scientific Reports, and IEEE Access suggests a tendency to publish in high-impact, multidisciplinary outlets spanning computer science, medicine, and engineering. Overall, the findings indicate that research on digital twins in healthcare is a globally collaborative effort, with major contributions from countries such as the USA, India, the United Kingdom, and Germany. The emphasis on keywords like digital twin, artificial intelligence, and machine learning underscores the technological advancements driving this field.

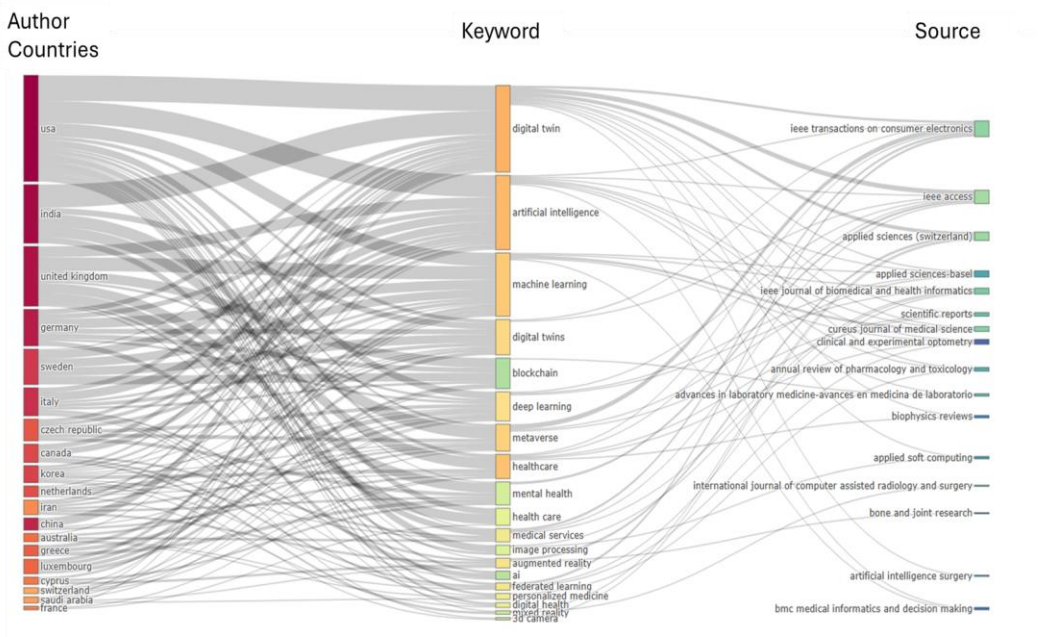


Figure 2. Analysis of the relationships between authors' countries, keywords and sources.

Analysis of the frequency of author keywords

Figure 3 presents a frequency analysis of the keywords used by authors in the 72 analyzed articles. The most frequently occurring keyword, "Digital Twin" appears in 27 articles, underscoring its central role in the research. "Artificial Intelligence" follows closely with 14 mentions, highlighting AI's pivotal role in the development and implementation of digital twins. The keyword "Healthcare" appears 12 times, reinforcing the study's domain. Additionally, terms such as "Machine Learning", "Metaverse", and "Deep Learning" reflect the technological foundations of digital

twin solutions. Meanwhile, "Augmented Reality" and "Federated Learning" suggest emerging applications and privacy-preserving approaches in the field. Notably, "Medical Services" appears only four times, indicating a gap in research on how DTs translate into practical healthcare solutions. This analysis reveals a strong focus on the technological aspects of digital twins, while the relatively lower frequency of "Medical Services" suggests the need for further exploration of their real-world applications in healthcare.

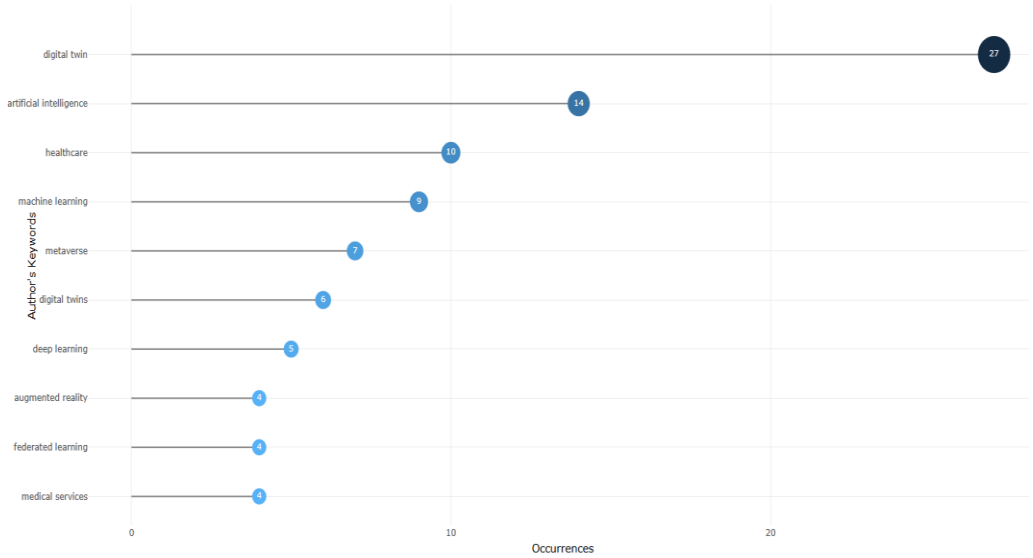


Figure 3. Keyword frequency analysis of the authors of the selected articles.

Factorial analysis results

Figure 4 presents a factorial analysis of bigrams extracted from the abstracts of 72 articles, revealing two dominant dimensions, Dim 1 and Dim 2, that together account for 61.64% of the variance in the data. Dim 1 (32.41%) captures the contrast between clinical applications and technological advancements. On one end, terms related to clinical decision-making, drug discovery, and research highlight the practical application of DTs in healthcare. On the other end, terms such as deep learning, neural networks, and proposed systems emphasize the technical foundations of DTs technology. Dim 2 (29.23%) differentiates between theoretical exploration and practical implementation. One pole, characterized by terms like “healthcare system” and “patient care”, underscores the real-world impact of digital twins on healthcare delivery. The opposite pole, with terms such as “potential applications” and “smart healthcare” suggests an ongoing discussion on future possibilities and advancements. The analysis identifies three distinct research clusters. The first, located in the top right quadrant of Figure 4, focuses on the clinical applications of DTs, with terms such as "clinical decision", "drug discovery", and "clinical practice" indicating an interest in how DTs can enhance patient care, from diagnosis and treatment planning to drug development. The second, situated in the left quadrant, highlights the technological advancements driving DTs development. Bigrams such as "deep

learning", "neural networks", and "proposed system" suggest a strong focus on artificial intelligence, machine learning, and other cutting-edge technologies that enhance DTs models for healthcare. The third cluster, positioned in the bottom right quadrant, emphasizes the integration of DTs within the healthcare system. Bigrams such as "healthcare system", "healthcare industry", and "patient care" suggest a focus on how DTs can be implemented and scaled within existing healthcare structures to improve efficiency, decision-making, and patient outcomes. Overall, the findings reveal a dynamic research landscape that balances theoretical exploration with practical implementation. The strong emphasis on clinical applications reflects a growing interest in translating DTs research into real-world patient care solutions. The focus on technological advancements highlights ongoing efforts to refine neural networks and deep learning techniques that support DTs development. Finally, the emphasis on healthcare system integration underscores the importance of seamless adoption within healthcare organizations. However, further exploration is needed to bridge the gap between technological innovation and its practical applications in healthcare.

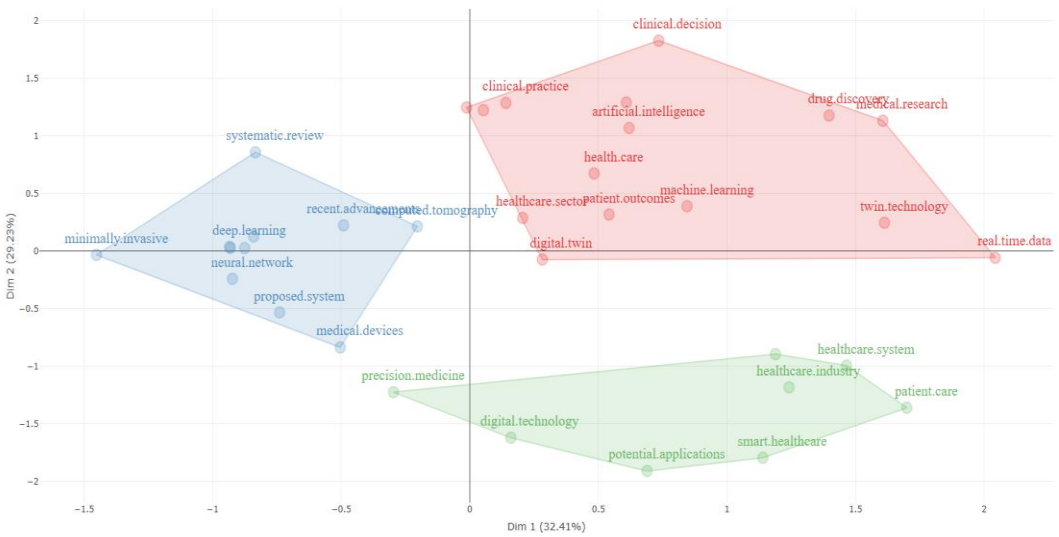


Figure 4. Factor analysis (MCA) of the abstract bi grams of the selected articles. The color of the points represents the cluster to which they belong.

Thematic analysis

Figure 5 presents a thematic map of the bigrams extracted from the abstracts of the 72 selected articles, revealing several distinct thematic clusters. A key cluster, located in the upper right quadrant, includes the bigrams “digital twin”, “artificial intelligence”, and “machine learning.” This cluster represents the motor (or core) themes of the field, highlighting areas of rapid technological advancement and active scholarly research. The presence of these concepts suggests their transformative potential in healthcare and their role at the forefront of innovation. Another significant group of core theme clusters, also in the upper right quadrant, consists of “clinical practice”, “personalized medicine”, “healthcare”, “augmented reality”, and

"digital technology". These motor themes serve as the foundational principles for DTs applications in healthcare, providing the structural basis upon which more advanced solutions are built. The only basic theme identify is "deep learning" (bottom right quadrant). Given that deep learning serves as the backbone for many cutting-edge technologies in this domain, further exploration of its applications and implications is essential. This theme will be discussed in greater detail in the next section. In the upper left quadrant, a niche theme emerges, encompassing "potential applications", "proposed systems", and "enabling technologies." This cluster aligns with previous analyses, reinforcing the notion that emerging technologies are still in the early stages of adoption. The literature is gradually engaging with these innovations, but widespread implementation remains limited. Finally, clusters related to "surgical robot", "complex medical", "recent advancements" and "computed tomography" are identified as emerging themes. These concepts represent highly promising yet specific applications of DTs within the broader healthcare landscape, indicating areas of ongoing exploration and future potential.

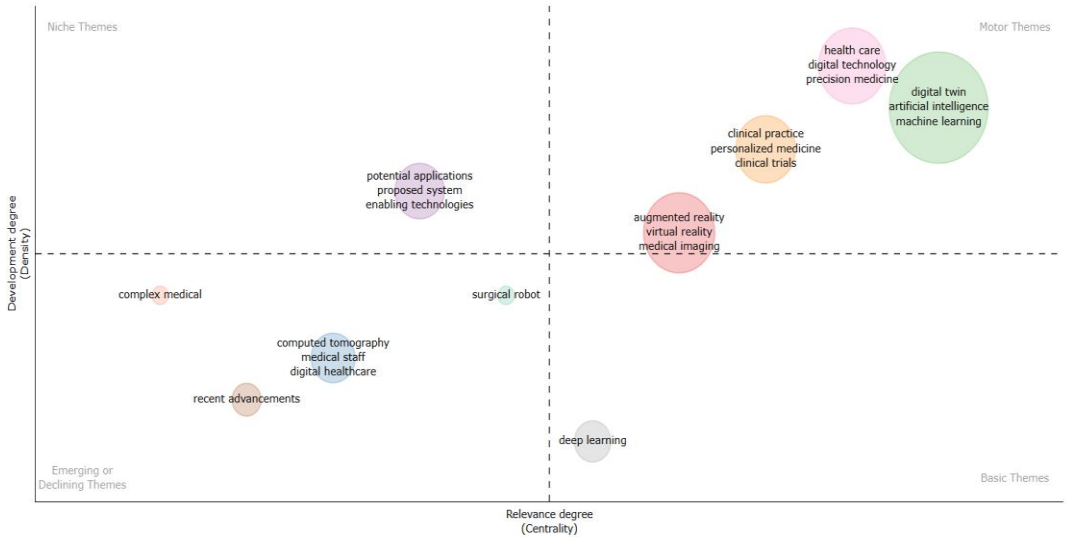


Figure 5. Thematic analysis of abstract bigrams. The chart is divided into 4 quadrants (starting from the top left quadrant and proceeding clockwise): niche themes, driving themes, basic themes and declining/emerging themes.

The bibliometric analysis presented highlights the rapid growth of interest in digital twins and the use of high technology in healthcare and the need for scientific maturation. Research efforts increasingly focus on AI-driven simulations, cross-disciplinary technologies, and real-time data integration, underscoring the expanding role of DTs across various domains. As this momentum builds, the Rome Technopole project positions itself at the cutting edge of these developments, particularly by advancing personalized surgical planning and enabling remote procedures with high-fidelity feedback systems.

Results from the SLR

As we presented in the previous Section, the application of DTs within the healthcare sector doesn't appear as an isolated technology, but they are often integrated with multiple types of other technologies, coming from different fields, such as medical engineering or computer science. We identified a diverse landscape of these technologies, categorized into several distinct but interconnected domains. Artificial intelligence (AI) and machine learning (ML) emerged as a foundational pillar, with applications spanning from image analysis using models like YOLOv3 and ResNet, (Zinchenko et al., 2021) to the development of explainable, human-centered, and trustworthy AI systems. Generative models such as ClinicalGAN (Chandra et al., 2024) and advanced learning paradigms like federated learning (Ali M. et al., 2023), split learning and FedAVG (Stephanie et al., 2024) were also noted, alongside specialized applications like neural radiance fields (NeRFs), Neuralangelo (Kleinbeck et al., 2024), and models for medical dialogue such as Med-PaLM (Vidovszky,et al., 2024) Furthermore, AI-powered digital health tools like Wysa and Ada Health (Abilkaiyrkyzy et al., 2024) were observed, indicating a trend towards personalized and accessible healthcare solutions. Virtual and augmented reality (VR/AR) technologies play a crucial role in creating immersive and interactive digital twin environments. AR and VR serve as pivotal tools for bridging the gap between the real world and its virtual counterpart. By enabling immersive visualization and interaction, VR and AR play a crucial role in enhancing the practical application of DTs in clinical settings, such as simulating surgeries for training or for planning or modelling complex physiological processes. Platforms like Unity 3D (Sunt et al., 2023; Balasubramanyam et al., 2024; Zackoff et al., 2023), coupled with hardware such as Meta Quest 2 (Balasubramanyam et al., 2024), Oculus Quest 2 (Zackoff,et al., 2023), HTC Vive Pro (Balasubramanyam et al., 2024), and HoloLens (Barcali et al., 2022; Seetohul et al., 2023; Aliani et al., 2024; Mikolajewski et al., 2024; Prasad et al., 2024; Balasubramanyam et al., 2024), are employed for diverse applications, including surgical training (e.g., Simbionix ArthroMentor, Simendo arthroscopy simulator) and visualization of complex anatomical structures (e.g., UCSF ChimeraX, YASARA). The use of haptic training and specialized VR/AR systems like VisAR (Seetohul et al., 2023), MetaMedicsVR (Hulsen et al. 2024), and Narupa iMD (Hulsen et al. 2024) further underscores the growing importance of these technologies in medical education and procedural planning. Biomedical imaging and diagnosis are significantly enhanced by DTs through the integration of advanced imaging modalities. Techniques such as shear-wave elastography (Bjelland et al., 2022), CBCT imaging (Lee et al., 2023), digital breast tomosynthesis (Pinto et al., 2023), and 3D echocardiography (Sachdeva et al., 2024) provide detailed anatomical and functional data that can be integrated into DTs models. Medical imaging equipment from manufacturers like Philips (e.g., IntelliVUE MX800) and GE Healthcare (e.g., Vivid S6), along with vein visualization technologies like AccuVein and NextVein (Seetohul et al., 2023), contribute to more precise diagnostics and treatment planning within the digital twin framework. Robotic surgery and medical planning represent another area where DTs are transforming healthcare. Surgical robots like the DaVinci system (Seetohul et al.,

2023), along with advanced planning software such as Virtual Cardiac Surgery Planning, ImmersiveView Surgical Plan, iPlan Flow, HeartFlow Analysis technology (Wu et al., 2022), Philips HeartNavigator, Acorys Mapping system and Feops' HEARTguide (Sun et al., 2023), enable surgeons to simulate and optimize procedures before execution. This integration of robotic systems with digital twin technology allows for greater precision, minimally invasive approaches, and improved patient outcomes. Biomedical simulation models form the core of many DTs applications. Tools and platforms such as BioSecure (Elkefi, et al., 2022), HumMod (Montgomery et al., 2023), Archimedes (Montgomery et al., 2023), UCSF ChimeraX (Hulsen et al. 2024), and the concept of the Digital Human Twin are used to simulate physiological processes, disease progression, and treatment responses. These models, often addressing specific surgical simulation challenges (e.g., EndoVis, SAR-RARP50, CATARACTS), provide valuable insights for personalized medicine and clinical decision-making. Information systems and digital health infrastructure are essential for the effective implementation of digital twins in healthcare. Electronic health records, personal wearables and remote monitoring devices, cloud-based personal health record systems, and platforms like Ali Health (Liu et al., 2019), Baidu Medical Cloud (Liu et al., 2019), Health@Hand (Elkefi et al., 2022), CloudDTH (Liu et al., 2019), eHealth systems (Liu et al., 2019), HospiTwin (Elkefi et al., 2022), and HealthVault (Liu et al., 2019) facilitate data collection, integration, and analysis within the digital twin environment. This interconnectedness promotes better communication between healthcare providers and patients, enabling more proactive and personalized care. Furthermore, 3D modeling and project tools like Autodesk Revit (Madubuike et al., 2023), Rhino (Wang et al., 2024), CAD modeling software, AnyLogic Simulink (Balasubramanyam, et al., 2024), COMSOL Multiphysics (Balasubramanyam et al., 2024), and Bentley Architecture (Madubuike et al., 2023) are used to create detailed representations of physical spaces and medical devices within the digital twin framework. Algorithms and networks, including the pendulum algorithm (Jiang et al., 2022), Levenberg-Marquardt algorithm (Sai et al., 2024), Damped Least-Squares algorithm (Sai et al., 2024), Time Sensitive Networking (Lu et al., 2023), and DetNet (Lu et al., 2023), provide computational grounds for simulating complex interactions and optimizing system performance. Specialized devices like building information models, DTs for 3D print clouds, VITASCOPE (Wang et al., 2024), Eclipse Ditto (Balasubramanyam et al., 2024), eMI MED, and MoodPath (Abilkaiyrkyzy et al., 2024) further contribute to the diverse applications of DTs in healthcare. Robot and human collaboration platforms like ManipulaTHOR (Long et al., 2023), iGibson (Long et al., 2023), ThreeDworld (Long et al., 2023), SAPIEN (Long et al., 2023), dVRL platform (Long et al., 2023), and AMBF platform (Long et al., 2023), along with technologies for education, metaverse, and recognition, including Virtual Classroom (Preshaw et al., 2024), DTCoach (Elkefi et al., 2022), Metaversespinal, MeTAI metaverse (Wang et al., 2022), face and posture emotion recognition using techniques like Haar classifiers, highlight emerging trends in training, communication, and personalized interventions within digital twin-enabled healthcare environments. Given the technologies presented, the use of haptics and

deep learning technologies (in particular PINNs) in the context of the Rome Technopole project relating to '*phygital*' is fully in the new growth technologies in use in this field, positioning itself at the existing technology frontier. Considering the wide world of medicine, however, it is necessary to investigate which clinical areas have been involved in DTs works in the past.

The distribution of research on DTs across various clinical fields reveals a concentration of interest in several key areas (Table 1). Medical education, surgery, and orthopedics emerge as prominent domains, each represented by a substantial number of articles (11, 10, and 10, respectively). In surgery, DTs emerge as transformative tools that enable preoperative planning, intraoperative guidance, and postoperative evaluation. By creating highly detailed virtual replicas of a patient's anatomy, DTs empower surgeons to simulate procedures, identify potential complications, and optimize surgical strategies, adapting them to specific cases. This capability not only enhances surgical precision but also reduces risks, shortens recovery times, and improves overall patient outcomes. In precision medicine, DTs provide a personalized approach to treatment by leveraging patient-specific data. These models integrate information from imaging, genomics, and other diagnostic tools to predict how an individual might respond to various treatments. This approach enables clinicians to adapt interventions to the unique characteristics of each patient. The distribution of clinical application suggests a strong focus on utilizing DTs for training purposes, surgical planning and simulation, and the management of musculoskeletal conditions. Precision medicine and preventive care also represent a significant area of investigation (9 articles), indicating a growing interest in leveraging digital twins for personalized healthcare strategies and proactive interventions. The use of DTs in precision medicine also includes disease management, such as monitoring disease progression and adjusting treatments to evolving situations. By offering interactive and immersive learning environments, DTs allow medical students and trainees to practice procedures, face complicated physiological systems, and visualize the effects of interventions in a risk-free setting. Several other clinical fields demonstrate a moderate level of research activity. Cardiovascular applications, diagnosis and treatment methodologies, oncology, and neurosurgery are each represented by a smaller but notable number of articles (5, 5, 4, and 4, respectively), highlighting the potential of DTs in addressing complex conditions and optimizing therapeutic approaches within these specialties. Pharmacy and drug discovery also feature in the literature (4 articles), suggesting the exploration of digital twins for accelerating drug development and personalized pharmaceutical interventions.

Specific medical disciplines such as rehabilitation, telemedicine, cardiology, and radiology are represented by a smaller number of studies (3 articles each), indicating emerging interest in these areas and potential for future expansion of digital twin applications. Other areas, such as pulmonology and dentistry, have two articles each. Finally, a selection of highly specialized clinical areas, including brain diseases, urology, gastrointestinal conditions, maxillofacial surgery and mental health, are each represented by a single article. While these areas currently have a limited number of publications related to digital twin technology, their inclusion suggests a

broadening scope of investigation and the potential for future growth as technology matures and its applications become more widely explored across diverse medical specialties.

Table 1. Distribution of research on digital twins across various clinical fields.

Clinical application		
Clinical field	Number of articles in this field	Reference(s)
Medical education	11	Balasubramanyam et al. (2024); Edgar et al. (2024); Hulsen (2024); Sai et al. (2024); Cellina et al. (2023); Kim & Kim (2023); Lee et al. (2023); Zackoff et al. (2023); Barcali et al. (2022); Denecke & Baudoin (2022); Zhang & Tai (2022)
Surgery	10	Aliani et al. (2024); Ding et al. (2024); Baumann et al. (2023); Long et al. (2023); Jiang et al. (2022); Razek (2023); Sun et al. (2023); Barcali et al. (2022); Denecke & Baudoin (2022); Zinchenko & Song, (2021)
Orthopaedics	11	Sun et al. (2023) ; Barcali et al. (2022); Bjelland et al. (2022); Lisacek-Kiosoglous et al. (2023); Seetohul et al. (2023); Liang et al. (2024); Ding et al. (2024); Aliani et al. (2024); Prasad et al. (2024); Zsidai et al. (2023); Zhou et al. (2024)
Precision medicine and preventive care	9	Sun et al. (2023); Vallée (2023); Suchetha et al. (2024); Sai et al. (2024); Bruynseels et al. (2018); Balasubramanyam et al. (2024); Liu et al. (2019) ;Venkatesh et al. (2024); Milne-Ives et al. (2022)
Cardiovascular	5	Sun et al. (2023); Wu et al. (2022); Ding et al. (2024); Aliani et al. (2024); Rouhollahi et al. (2023)
Diagnosis and treatment	5	Sun et al. (2023); Pregowska & Perkins (2024); Balasubramanyam et al. (2024); Venkatesh et al. (2024); Sharma et al. (2024)
Pharmacy and drug discovery	4	Sun et al. (2023); Balasubramanyam et al. (2024); Cellina et al. (2023); Venkatesh et al. (2024)

Oncology	4	Barcali et al. (2022); Wu et al. (2022); Aliani et al. (2024); Prasad et al. (2024)
Neurosurgery and neuroscience	4	Barcali et al. (2022); Seetohul et al. (2023); Prasad et al. (2024); Fekonja et al. (2024)
Rehabilitation	3	Denecke & Baudoin (2022); Mikolajewski et al. (2024); Tao et al. (2024)
Telemedicine	3	Denecke & Baudoin (2022); Kim & Kim (2023); Hulsen (2024)
Cardiology	3	Seetohul et al. (2023); Mikolajewski et al. (2024); Sachdeva et al. (2024)
Radiology	3	Pesapane et al. (2022); Geissler et al. (2021); Panayides et al. (2020)
Pneumology	2	Zhang & Tai (2022); Montgomery et al. (2023)
Dentistry	2	Lee et al. (2023); Preshaw et al. (2024)
Brain diseases	1	Wu et al. (2022)
Urology	1	Kim & Kim (2023)
Gastrointestinal	1	Seetohul et al. (2023)
Maxillofacial surgery	1	Aliani et al. (2024)
Mental health	1	Abilkaiyrkyzy et al. (2024)

Despite the various instruments and clinical applications in literature, considering how sensitive the medical field is, it is important to evaluate the advantages and disadvantages identified. This is necessary to assess the benefits and costs of using frontier technologies and DTs in a field where patients' lives are at stake. This step is critical before we can have full deployment of these technologies. Disadvantages and advantages identified in the literature are presented in Table 2 (advantages) and Table 3 (disadvantages). The reviewed literature highlights a range of advantages associated with the application of digital twin technology in healthcare. The most frequently cited benefit pertains to improvements in patient care (20 articles), encompassing enhancements in both pre-clinical and post-clinical phases, and facilitating more personalized treatments. This broad category is supported by numerous studies (e.g., Lisacek-Kiosoglous et al., 2023; Kim & Kim, 2023; Zinchenko & Song, 2021; Chandra et al., 2024), indicating a strong consensus on the potential of digital twins to revolutionize patient management. Several other key advantages emerged prominently. Digital twins were frequently reported to improve physicians' accuracy in surgery and decision-making (10 articles), with studies such as Wu et al. (2022), Bjelland et al. (2022) and Lu et al. (2023), providing evidence for this claim. Similarly, the technology's potential to enhance medical education through flexible and adaptable online learning was highlighted in 10 articles (e.g., Long et al., 2023; Preshaw et al., 2024). Real-time data extraction, precise treatments, and improved predictive abilities were each identified as advantages in

nine articles. Studies such as Sun et al. (2023), Lee et al. (2023), and Pinto et al. (2023) support the role of digital twins in enabling timely data analysis, tailoring interventions to individual patient needs, and forecasting disease progression or treatment outcomes. Improvements in surgery, including applications with robotic systems, were noted in eight articles (e.g., Yang, 2023; Wang, 2024; Vallée, 2023), further emphasizing the technology's impact on surgical practice. Enhancements in diagnosis were reported in six articles (e.g., Zhang & Tai, 2022, Yang, 2023, Bhattad & Jain, 2020), while interoperability and improvements to healthcare structures were each mentioned in five articles (e.g., Bjelland et al., 2022; Yang, 2023; Wang, 2024). Real-time monitoring was identified as a benefit in three articles (e.g., Liu et al. 2019; Venkatesh et al., 2024), and improvements in security and increasing drug development were each noted in two articles (e.g., Upreti et al., 2024; Hulsen, 2024). Finally, several more specific advantages were each mentioned in a single article: the use of finite element (FE) methods for non-invasive, controllable, and repeatable procedures (Sun et al., 2023), efficiency of visualization and reduction of exposure to ionizing radiation (Barcali et al., 2022), and the establishment of a link between the real and virtual worlds (Garg et al., 2022). This distribution of reported advantages underscores the multifaceted impact of digital twins technology across various aspects of healthcare, from patient care and surgical precision to medical education and drug development. The concentration of articles on patient care, surgical accuracy, and medical education suggests these areas are currently the primary focus of research and application, while the presence of more specific advantages indicates the potential for further exploration and development in diverse sub-domains. At the same time, several disadvantages associated with digital twin technology in healthcare were identified in the reviewed literature. Security and privacy concerns emerged as the most frequently cited drawback, mentioned in 16 articles (e.g., Zhang & Tai, 2022; Denecke & Baudoin, 2022; Stephanie et al., 2024). This highlights the critical need for robust data protection measures and ethical considerations surrounding the sensitive information managed within digital twin systems. The scarcity, accuracy, and quality of data were identified as a significant challenge in 10 articles (e.g., Sun et al., 2023, Wu et al., 2022; Geissler et al., 2021). This underscores the importance of reliable data sources and rigorous data validation processes to ensure the integrity and effectiveness of digital twin models. Ethical, social, and legal risks were also frequently discussed, appearing in seven articles (e.g., Sun et al., 2023, Pregowska & Perkins, 2024; Vidovszky et al., 2024), emphasizing the need for careful consideration of the broader societal implications of this technology. High costs associated with implementation and maintenance were noted in four articles (e.g., Bjelland et al., 2022; Lisacek-Kiosoglous et al., 2023) highlighting the economic barriers that may hinder widespread adoption. The dependency on the accuracy of simulations and potential model errors was identified as a disadvantage in three articles (Sun et al., 2023; Barcali et al., 2022; Kim & Kim, 2023), emphasizing the importance of continuous model refinement and validation. Medical interoperability, referring to the ability of different systems and devices to exchange and utilize data, was also mentioned in three articles (e.g., Yang, 2023; Ding et al., 2024; Khater et al., 2024). Several disadvantages were noted in two

articles each: the need for further validation of digital twin models (e.g., Sun et al., 2023; Wu et al., 2022), challenges in establishing accurate and intuitive action mapping between human input devices and surgical robots (e.g., Long et al., 2023; Jiang et al., 2022), the potential for incorrect connections between the real and virtual worlds due to sensor reliability and accuracy (e.g., Seetohul et al., 2023; Liu et al., 2019), and the presence of biases in data or model design (e.g., Gwon et al., 2024, Pesapane et al., 2022). A range of highly specific disadvantages were each mentioned in a single article: vergence-accommodation conflict in VR/AR applications (Barcali et al., 2022), challenges in connecting phenomena at different scales and calibrating model parameters (Wu et al., 2022), technological limitations related to computational techniques, model selection, validation, uncertainty quantification, and data interoperability (Zhang & Tai, 2022), dependence on Magnetic Resonance Imaging (MRI) (Bjelland et al., 2022), discomfort associated with wearable devices (Kim & Kim, 2023), the difficulty of building realistic physical simulations with high-fidelity scene visualization (Long et al., 2023) the influence of ambient air humidity during the setting phase (a phenomenon that is not always easily simulated, Lee et al., 2023), inconsistency in delivery and assessment methods across online learning platforms (Preshaw et al., 2024), the need for advanced haptic training tools (which has a huge impact in the diffusion and in the costs of implementation of the different solutions available, Preshaw et al., 2024), concerns related to distribution through virtual pharmacies (Yang, 2023), challenges in real-time modeling of tissues (especially computational problems, Razek, 2023), limitations in real-time information updates and bi-directional coordination in hospital facilities management (Madubuike et al., 2023), and the potential lack of individualization in certain applications (Milne-Ives et al., 2022). This diverse array of disadvantages highlights the ongoing challenges and areas for improvement in the development and implementation of digital twin technology in healthcare. The prevalence of concerns related to security, data quality, and ethical considerations underscores the need for careful planning and robust safeguards to ensure responsible and effective utilization of this technology. Of the various advantages and disadvantages identified, the Rome Technopole project is at the forefront in several respects. Firstly, the development of real-time soft tissue simulation systems using computational capacity reduction techniques required (using PINNs) makes it possible to solve one of the problems listed above. Other projects in the Rome Technopole are currently working on solving another of the problems listed above, namely that of security and data. In the future, the implementation of phygital technology within the project will have to take account of what has been identified, especially in relation to the communication and sensor issues adopted to avoid the serious problem of poor data quality.

Table 2. Advantages identified in the literature on the application of DTs and emerging technologies in the medical field.

Advantages	*Number of articles*	*Reference(s)*
Patient care (improving pre-clinical phase and post-clinical phase, reducing the distance for personalised treatments)	20	Lisacek-Kiosoglous et al. (2023); Kim & Kim (2023); Zinchenko & Song, (2021); Preshaw et al. (2024); Yang (2023); Gwon et al. (2024); Liang et al. (2024); Jiang et al. (2022) ; Upreti et al. (2024); Mikolajewski et al. (2024); Razek (2023); Suchetha et al. (2024); Sai et al. (2024); Cellina et al. (2023); Elkefi & Asan (2022); Subramanian et al. (2022); Hulsen (2024); Tao et al. (2024); Stephanie et al. (2024); Chandra et al. (2024)
Improves the physicians' accuracy in surgery and decision making	10	Wu et al. (2022); Bjelland et al. (2022); Liang et al. (2024); Balasubramanyam et al. (2024); Cellina et al. (2023); Elkefi & Asan (2022); Montgomery et al. (2023); Kleinbeck et al. (2024); Khater et al. (2024); Lu et al. (2023)
Improves medical education (Flexibility and adaptability of online learning)	10	Long et al. (2023); Preshaw et al. (2024); Yang (2023); Pregowska & Perkins (2024); Aliani et al. (2024); Edgar et al. (2024); Mikolajewski et al. (2024); Razek (2023); Vallée (2023); Jamshidi et al. (2023)
Real-time data extraction	10	Abilkaiyrkyzy et al. (2024); Ding et al. (2024); Sun et al. (2023); Zhang & Tai (2022); Denecke & Baudoin (2022)Lisacek-Kiosoglous et al. (2023); Wang (2024); Garg et al. (2022); Bhatia (2024); Joo et al. (2024)
Precise treatments	9	Sun et al. (2023); Lee et al. (2023) Seetohul et al. (2023); Milne-Ives et al. (2022); Kim & Kim (2023); Elkefi & Asan (2022); Chandra et al. (2024); Khater et al. (2024); Bhattad & Jain (2020)

Improves predicting ability	9	Lisacek-Kiosoglous et al. (2023);Upreti et al. (2024); Zsidai et al. (2023); Panayides et al. (2020); Vallée (2023); Pesapane et al. (2022); Khater et al. (2024); Kulkarni et al. (2024); Pinto et al. (2023)
Improves surgery	8	Yang (2023); Liang et al. (2024); Jiang et al. (2022); Baumann et al. (2023); Vallée (2023); Ding et al. (2024); Wang (2024); Khater et al. (2024)
Improves diagnosis	6	Zhang & Tai (2022); Yang (2023); Sharma et al. (2024); Vidovszky et al. (2024); Sachdeva et al. (2024); Bhattad & Jain (2020)
Interoperability	5	Bjelland et al. (2022); Yang (2023); Ding et al. (2024); Prasad et al. (2024); Balasubramanyam et al. (2024)
Improves healthcare structures	5	Wang (2024); Lisacek-Kiosoglous et al. (2023); Yang (2023); Madubuike (2023); Vidovszky et al. (2024)
Real time monitoring	3	Liu et al. (2019); Venkatesh et al. (2024); Bhatia (2024)
Improves security	2	Upreti et al. (2024); Bhatia (2024)
Increasing drug development	2	Vidovszky et al. (2024); Hulsen (2024)
non-invasiveness and repeatability	1	Sun et al. (2023)
Efficiency of visualization	1	Barcali et al. (2022)
Reduction of exposure to ionizing radiation	1	Barcali et al. (2022)
Link between real world and virtual world	1	Garg et al. (2022)

Table 3. Disadvantages identified in the literature on the application of DTs and emerging technologies in the medical field.

Disadvantages	*Number of articles*	*Reference(s)*
Security and privacy	16	Zhang & Tai (2022); Denecke & Baudoin (2022); Preshaw et al. (2024) Gwon et al. (2024); Upreti et al. (2024); Khater et al. (2024); Stephanie et al. (2024); Suchetha et al. (2024); Bruynseels et al. (2018); Balasubramanyam et al. (2024); Cellina et al. (2023); Venkatesh et al. (2024); Hulsen (2024); Garg et al. (2022); Ali et al. (2023); Wang (2024)
Scarsity, accuracy and quality of data	10	Sun et al. (2023); Wu et al. (2022); Upreti et al. (2024); Baumann et al. (2023); Pesapane et al. (2022); Cellina et al. (2023); Milne-Ives et al. (2022); Rouhollahi et al. (2023); Geissler et al. (2021); Vidovszky et al. (2024)
Ethical, social and legal risks	9	Sun et al. (2023); Pregowska & Perkins (2024); Upreti et al. (2024); Suchetha et al. (2024); Bruynseels et al. (2018); Balasubramanyam et al. (2024); Vidovszky et al. (2024); Zhou et al. (2025); Joo et al. (2024)
High costs	4	Bjelland et al. (2022); Lisacek-Kiosoglous et al. (2023); Liang et al. (2024); Lu et al. (2023)
Dependency from accuracy of simulation and model's errors	3	Sun et al. (2023); Barcali et al. (2022); Kim & Kim (2023)
Medical interoperability	3	Yang (2023); Ding et al. (2024); Khater et al. (2024)
Need for further validation	2	Sun et al. (2023); Wu et al. (2022)
Scarse mechanism between human input device and surgical robots	2	Long et al. (2023); Jiang et al. (2022)
Incorrect connection between (sensor reliability and accuracy)	2	Seetohul et al. (2023); Liu et al. (2019)
Biases	2	Gwon et al. (2024); Pesapane et al. (2022)

Vergence-accomodation conflict	1	Barcali et al. (2022)
Scales and calibration of model parameters	1	Wu et al. (2022)
Technological limitations	1	Zhang & Tai (2022)
Dependence on MRI	1	Bjelland et al. (2022)
Discomfort of the wearable devices	1	Kim & Kim (2023)
Complexity in realistically and faithfully simulating physical interactions	1	Long et al. (2023)
Humidity of the ambient air during the setting phase	1	Lee et al. (2023)
Inconsistency in delivery and assessment methods across online platforms	1	Preshaw et al. (2024)
Haptic training required	1	Preshaw et al. (2024)
Virtual Pharmacies using Extended Reality are complex to implement	1	Yang (2023)
Real time modeling of tissues	1	Razek (2023)
Lack of real time information	1	Madubuike et al. (2023)
Lack of individualization	1	Milne-Ives et al. (2022)

Preliminary Conclusions

Looking ahead, the future of DTs in healthcare lies in broader applications of these technologies. Personalized medicine, adaptive diagnostics, and real-time surgical interventions are among the key areas of expansion. The Rome Technopole project exemplifies this forward-thinking vision by emphasizing not only immediate training applications but also the potential for remote and augmented surgical systems to redefine medical practices. Through its innovative use of AI, haptics, and immersive technologies, the project has already made significant strides. Its publications, including studies on haptic interactions with virtual deformable objects, reflect this progress. Indeed, this project, drawing upon existing literature and through the integration of various haptic technologies and VR technologies, is moving towards highly evolved phygital twins technology, also known as autonomous twins (Zhang et al., 2024). Autonomous twins operate independently while seamlessly interacting with the physical world, potentially creating metaverses populated by autonomous virtual entities. This evolution promises to revolutionize healthcare through applications such as autonomic DTs brains for personalized interventions, realistic surgical training with tailored feedback, and ultimately, the realization of precision medicine by accelerating medical discoveries and improving treatment outcomes. As digital twins evolve towards this autonomous stage, their integration into healthcare will undoubtedly lead to new standards of precision, efficiency, and accessibility, paving the way for revolutionary advancements in medical care.

Acknowledgments

This paper is done within the Research Project "Phygital Twin Technologies for Innovative Surgical Training & Planning (ECS 0000024 – Rome Technopole)" within the Rome Technopole, which is the first multi-technology hub for education, research and technology transfer in the fields of Energy Transition and Sustainability, Digital Transformation and Health & Bio-Farma. The project is part of the National Recovery and Resilience Plan (PNRR). It creates for the first time an innovation ecosystem for the Lazio Region, in which 7 universities, 4 research organizations, the Lazio Region the Municipality of Rome and other public bodies, 20 industrial groups and enterprises participate.

References

Abilkaiyrkyzy, A., Laamarti, F., Hamdi, M., & Saddik, A. E. (2024). Dialogue System for Early Mental Illness Detection: Toward a Digital Twin Solution. *IEEE ACCESS*, *12*, 2007–2024.

Ali, M., Naeem, F., Tariq, M., & Kaddoum, G. (2023). Federated Learning for Privacy Preservation in Smart Healthcare Systems: A Comprehensive Survey. *IEEE Journal of Biomedical and Health Informatics*, *27*(2), 778–789.

Aliani, C., Morelli, A., Rossi, E., Lombardi, S., Civale, V. Y., Sardini, V., Verdino, F., et al. (2024). Realistic Texture Mapping of 3D Medical Models Using RGBD Camera for Mixed Reality Applications. *Applied Sciences (Switzerland)*, *14*(10). Multidisciplinary Digital Publishing Institute (MDPI).

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. Elsevier.

Avenali, A., Daraio, C., Di Leo, S., Matteucci, G., & Nepomuceno, T. (2023). Systematic reviews as a metaknowledge tool: caveats and a review of available options. *International Transactions in Operational Research*, *30*(6), 2761–2806. John Wiley & Sons, Ltd.

Aziz, S., Jung, D. W., Uz Zaman, U. K., & Aqeel, A. Bin. (2024). Digital Twins in Smart Manufacturing. *Handbook of Manufacturing Systems and Design: An Industry 4.0 Perspective* (pp. 53–68). Taylor & Francis.

Bagaria, N., Laamarti, F., Badawi, H. F., Albraikan, A., Velazquez, R. A. M., & El Saddik, A. (2020). Health 4.0: Digital Twins for Health and Well-Being. *Connected Health in Smart Cities* (pp. 143–152). Springer, Cham.

Balasubramanyam, A., Ramesh, R., Sudheer, R., & Honnavalli, P. B. (2024). Revolutionizing Healthcare: A Review Unveiling the Transformative Power of Digital Twins. *IEEE ACCESS*, *12*, 69652–69676.

Barcali, E., Iadanza, E., Manetti, L., Francia, P., Nardi, C., & Bocchi, L. (2022). Augmented Reality in Surgery: A Scoping Review. *Applied Sciences (Switzerland)*, *12*(14). MDPI.

Baumann, O., Lenz, A., Hartl, J., Bernhard, L., & Knoll, A. C. (2023). Intuitive teaching of medical device operation to clinical assistance robots. *International Journal of Computer Assisted Radiology and Surgery*, *18*(5), 865–870. Springer Science and Business Media Deutschland GmbH.

Bhatia, M. (2024). An AI-enabled secure framework for enhanced elder healthcare. *ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*, *131*.

Bhattad, P. B., & Jain, V. (2020). Artificial Intelligence in Modern Medicine - The Evolving Necessity of the Present and Role in Transforming the Future of Medical Care. *CUREUS JOURNAL OF MEDICAL SCIENCE*, *12*(5).

Bjelland, O., Rasheed, B., Schaathun, H. G., Pedersen, M. D., Steinert, M., Hellevik, A. I.,

& Bye, R. T. (2022). Toward a Digital Twin for Arthroscopic Knee Surgery: A Systematic Review. *IEEE Access*, *10*, 45029–45052. Institute of Electrical and Electronics Engineers Inc.

Broadus, R. N. (1987). Toward a definition of "bibliometrics." *Scientometrics*, *12*(5–6), 373–379. Kluwer Academic Publishers.

Bruynseels, K., de Sio, F. S., & van den Hoven, J. (2018). Digital Twins in Health Care: Ethical Implications of an Emerging Engineering Paradigm. *FRONTIERS IN GENETICS*, *9*.

Cellina, M., Cè, M., Alì, M., Irmici, G., Ibba, S., Caloro, E., Fazzini, D., et al. (2023). Digital Twins: The New Frontier for Personalized Medicine? *APPLIED SCIENCES-BASEL*, *13*(13).

Chandra, S., Prakash, P. K. S., Samanta, S., & Chilukuri, S. (2024). ClinicalGAN: powering patient monitoring in clinical trials with patient digital twins. *SCIENTIFIC REPORTS*, *14*(1).

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2011). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*, *5*(1), 146–166. Elsevier.

Denecke, K., & Baudoin, C. R. (2022). A Review of Artificial Intelligence and Robotics in Transformed Health Ecosystems. *Frontiers in Medicine*, *9*. Frontiers Media S.A.

De Santis, E., Le Jeune, Y., Marchal, M., Pacchierotti, C., & Vendittelli, M. (2024, October). Haptic interaction with virtual deformable objects. In *Proc. Italian Institute of Robotics and Intelligent Machines* (I-RIM) 3D 2024.

Ding, H., Seenivasan, L., Killeen, B. D., Cho, S. M., & Unberath, M. (2024). Digital twins as a unifying framework for surgical data science: the enabling role of geometric scene understanding. *Artificial Intelligence Surgery*, *4*(3), 109–138. OAE Publishing Inc.

Distefano, L., Fumagalli, A., Giampietro, G., Naffati, J., Romaniello, C., Vitale, V. M., & Vendittelli, M. (2023, ottobre 20). An Integrated Environment for Medical Training and Procedure Planning in Virtual Reality. https://doi.org/10.5281/zenodo.10722622

van Dinter, R., Tekinerdogan, B., & Catal, C. (2022). Predictive maintenance using digital twins: A systematic literature review. *Information and Software Technology*, *151*, 107008. Elsevier.

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. Elsevier.

Edgar, A. K., Chong, L. X., Wood-Bradley, R., Armitage, J. A., Narayanan, A., & Macfarlane, S. (2024). The role of extended reality in optometry education: a narrative review. *Clinical and Experimental Optometry*. Taylor and Francis Ltd.

El-Agamy, R. F., Sayed, H. A., AL Akhatatneh, A. M., Aljohani, M., & Elhosseini, M. (2024). Comprehensive analysis of digital twins in smart cities: a 4200-paper bibliometric study. *Artificial Intelligence Review*, *57*(6), 154. Springer.

Elkefi, S., & Asan, O. (2022). Digital Twins for Managing Health Care Systems: Rapid Literature Review. *JOURNAL OF MEDICAL INTERNET RESEARCH*, *24*(8).

Evangeline, P. (2020). Digital twin technology for "smart manufacturing." *Advances in computers* (Vol. 117, pp. 35–49). Elsevier.

Fekonja, L. S., Schenk, R., Schröder, E., Tomasello, R., Tomšič, S., & Picht, T. (2024). The digital twin in neuroscience: from theory to tailored therapy. *Frontiers in Neuroscience*, *18*, 1454856. Frontiers Media SA.

Garg, H., Sharma, B., Shekhar, S., & Agarwal, R. (2022). Spoofing detection system for e-

health digital twin using EfficientNet Convolution Neural Network. *MULTIMEDIA TOOLS AND APPLICATIONS*, *81*(19), 26873–26888.

Geissler, F., Heiss, R., Kopp, M., Wiesmüller, M., Saake, M., Wuest, W., Wimmer, A., et al. (2021). Personalized computed tomography - Automated estimation of height and weight of a simulated digital twin using a 3D camera and artificial intelligence. *ROFO-FORTSCHRITTE AUF DEM GEBIET DER RONTGENSTRAHLEN UND DER BILDGEBENDEN VERFAHREN*, *193*(4), 437–445.

Greenacre, M. (2017). Correspondence analysis in practice, third edition. *Correspondence Analysis in Practice, Third Edition*, 1–310. CRC Press.

Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. *Transdisciplinary perspectives on complex systems: New findings and approaches* (pp. 85–113). Springer.

Gwon, Y. N., Kim, J. H., Chung, H. S., Jung, E. J., Chun, J., Lee, S., & Shim, S. R. (2024). The Use of Generative AI for Scientific Literature Searches for Systematic Reviews: ChatGPT and Microsoft Bing AI Performance Evaluation. *JMIR Medical Informatics*, *12*. JMIR Publications Inc.

Hulsen, T. (2024). Applications of the metaverse in medicine and healthcare. *ADVANCES IN LABORATORY MEDICINE-AVANCES EN MEDICINA DE LABORATORIO*, *5*(2), 159–165.

Jamshidi, M., Sargolzaei, S., Foorginezhad, S., & Moztarzadeh, O. (2023). Metaverse and microorganism digital twins: A deep transfer learning approach. *APPLIED SOFT COMPUTING*, *147*.

Jiang, F., Jia, R., Jiang, X., Cao, F., Lei, T., & Luo, L. (2022). Human-Machine Interaction Methods for Minimally Invasive Surgical Robotic Arms. *Computational Intelligence and Neuroscience*, *2022*. Hindawi Limited

Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the Digital Twin: A systematic literature review. *CIRP journal of manufacturing science and technology*, *29*, 36–52. Elsevier.

Joo, Y., Camacho, D., Boi, B., Esposito, C., & Choi, C. (2024). Blockchain and Federated Learning Empowered Digital Twin for Effective Healthcare. *Human-centric Computing and Information Sciences*, *14*. Korea Information Processing Society.

Katsoulakis, E., Wang, Q., Wu, H., Shahriyari, L., Fletcher, R., Liu, J., ... & Deng, J. (2024). Digital twins for health: a scoping review. *NPJ digital medicine*, 7(1), 77.

Khater, H. M., Sallabi, F., Serhani, M. A., Barka, E., Shuaib, K., Tariq, A., & Khayat, M. (2024). Empowering Healthcare with Cyber-Physical System - A Systematic Literature Review. *IEEE Access*, *12*, 83952–83993. Institute of Electrical and Electronics Engineers Inc

Kim, E. J., & Kim, J. Y. (2023). The Metaverse for Healthcare: Trends, Applications, and Future Directions of Digital Therapeutics for Urology. *International Neurourology Journal*, *27*, S3–S12. Korean Continence Society.

Kleinbeck, C., Zhang, H., Killeen, B. D., Roth, D., & Unberath, M. (2024). Neural digital twins: reconstructing complex medical environments for spatial planning in virtual reality. *International Journal of Computer Assisted Radiology and Surgery*, *19*(7), 1301–1312. Springer Science and Business Media Deutschland GmbH.

Kulkarni, C., Quraishi, A., Raparthi, M., Shabaz, M., Khan, M. A., Varma, R. A., Keshta, I., et al. (2024). Hybrid disease prediction approach leveraging digital twin and metaverse technologies for health consumer. *BMC MEDICAL INFORMATICS AND DECISION MAKING*, *24*(1).

Lee, J.-H., Lee, H.-L., Park, I.-Y., On, S.-W., Byun, S.-H., & Yang, B.-E. (2023).

Effectiveness of creating digital twins with different digital dentition models and cone-beam computed tomography. *Scientific Reports*, *13*(1). Nature Research.

Liang, W., Zhou, C., Bai, J., Zhang, H., Jiang, B., Wang, J., Fu, L., et al. (2024). Current advancements in therapeutic approaches in orthopedic surgery: a review of recent trends. *Frontiers in Bioengineering and Biotechnology*, *12*. Frontiers Media SA.

Lisacek-Kiosoglous, A. B., Powling, A. S., Fontalis, A., Gabr, A., Mazomenos, E., & Haddad, F. S. (2023). Artificial intelligence in orthopaedic surgery EXPLORING ITS APPLICATIONS, LIMITATIONS, AND FUTURE DIRECTION. *Bone and Joint Research*, *12*(7), 47–454. British Editorial Society of Bone and Joint Surgery.

Liu, Y., Zhang, L., Yang, Y., Zhou, L. F., Ren, L., Wang, F., Liu, R., et al. (2019). A Novel Cloud-Based Framework for the Elderly Healthcare Services Using Digital Twin. *IEEE ACCESS*, *7*, 49088–49101.

Long, Y., Wei, W., Huang, T., Wang, Y., & Dou, Q. (2023). Human-in-the-Loop Embodied Intelligence With Interactive Simulation Environment for Surgical Robot Learning. *IEEE Robotics and Automation Letters*, *8*(8), 4441–4448. Institute of Electrical and Electronics Engineers Inc.

Lu, Y. Z., Zhao, G. F., Chakraborty, C., Xu, C., Yang, L., & Yu, K. P. (2023). Time-Sensitive Networking-Driven Deterministic Low-Latency Communication for Real-Time Telemedicine and e-Health Services. *IEEE TRANSACTIONS ON CONSUMER ELECTRONICS*, *69*(4), 734–744.

Madubuike, O. C., Anumba, C. J., & Agapaki, E. (2023). Scenarios for digital twin deployment in healthcare facilities management. *JOURNAL OF FACILITIES MANAGEMENT*.

Mikolajewski, D., Bryniarska, A., Wilczek, P. M., Myslicka, M., Sudol, A., Tenczynski, D., Kostro, M., et al. (2024). THE MOST CURRENT SOLUTIONS USING VIRTUAL-REALITY-BASED METHODS IN CARDIAC SURGERY – A SURVEY. *Computer Science*, *25*(1), 107–128. AGH University of Science and Technology Press.

Milne-Ives, M., Fraser, L. K., Khan, A., Walker, D., van Velthoven, M. H., May, J., Wolfe, I., et al. (2022). Life Course Digital Twins-Intelligent Monitoring for Early and Continuous Intervention and Prevention (LifeTIME): Proposal for a Retrospective Cohort Study. *JMIR RESEARCH PROTOCOLS*, *11*(5).

Montgomery, A. J., Litell, J., Dang, J., Flurin, L., Gajic, O., & Lal, A. (2023). Gaining consensus on expert rule statements for acute respiratory failure digital twin patient model in intensive care unit using a Delphi method. *BIOMOLECULES AND BIOMEDICINE*, *23*(6), 1108–1117.

Mourtzis, D., Angelopoulos, J., & Panopoulos, N. (2023). The future of the human–machine interface (HMI) in society 5.0. *Future Internet*, *15*(5), 162. MDPI.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, *372*. British Medical Journal Publishing Group.

Panayides, A. S., Amini, A., Filipovic, N., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D. J., et al. (2020). AI in Medical Imaging Informatics: Current Challenges and Future Directions. *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, *24*(7), 1837–1857.

Pesapane, F., Rotili, A., Penco, S., Nicosia, L., & Cassano, E. (2022). Digital Twins in Radiology. *JOURNAL OF CLINICAL MEDICINE*, *11*(21).

Pinto, M. C., Mauter, F., Michielsen, K., Biniazan, R., Kappler, S., & Sechopoulos, I. (2023). A deep learning approach to estimate x-ray scatter in digital breast tomosynthesis: From phantom models to clinical applications. *MEDICAL PHYSICS*, *50*(8), 4744–4757.

Pires, F., Cachada, A., Barbosa, J., Moreira, A. P., & Leitão, P. (2019). Digital twin in industry 4.0: Technologies, applications and challenges. *2019 IEEE 17th international conference on industrial informatics (INDIN)* (Vol. 1, pp. 721–726). IEEE.

Prasad, K., Fassler, C., Miller, A., Aweeda, M., Pruthi, S., Fusco, J. C., Daniel, B., et al. (2024). More than meets the eye: Augmented reality in surgical oncology. *Journal of Surgical Oncology*, *130*(3), 405–418. John Wiley and Sons Inc.

Pregowska, A., & Perkins, M. (2024). Artificial intelligence in medical education: Typologies and ethical approaches. *Ethics and Bioethics (in Central Europe)*, *14*(1–2), 96–113. Sciendo.

Preshaw, P. M., Ramseier, C. A., Loos, B. G., Balčiūnaitė, A., Crnić, T., Davey, K., Dommisch, H., et al. (2024). Contemporary educational methods in periodontology. *Journal of Clinical Periodontology*, *51*(S27), 117–192. John Wiley and Sons Inc.

Razek, A. (2023). Image-Guided Surgical and Pharmacotherapeutic Routines as Part of Diligent Medical Treatment. *Applied Sciences (Switzerland)*, *13*(24). Multidisciplinary Digital Publishing Institute (MDPI).

Rouhollahi, A., Willi, J. N., Haltmeier, S., Mehrtash, A., Straughan, R., Javadikasgari, H., Brown, J., et al. (2023). CardioVision: A fully automated deep learning package for medical image segmentation and reconstruction generating digital twins for patients with aortic stenosis. *COMPUTERIZED MEDICAL IMAGING AND GRAPHICS*, *109*.

Sachdeva, R., Armstrong, A. K., Arnaout, R., Grosse-Wortmann, L., Han, B. K., Mertens, L., Moore, R. A., et al. (2024). Novel Techniques in Imaging Congenital Heart Disease *JACC* Scientific Statement. *JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY*, *83*(1), 63–81.

Sai, S., Prasad, M., Garg, A., & Chamola, V. (2024). Synergizing Digital Twins and Metaverse for Consumer Health: A Case Study Approach. *IEEE TRANSACTIONS ON CONSUMER ELECTRONICS*, *70*(1), 2137–2144.

Seetohul, J., Shafiee, M., & Sirlantzis, K. (2023). Augmented Reality (AR) for Surgical Robotic and Autonomous Systems: State of the Art, Challenges, and Solutions. *Sensors*, *23*(13). Multidisciplinary Digital Publishing Institute (MDPI).

Sharma, V., Kumar, A., & Sharma, K. (2024). Digital twin application in women's health: Cervical cancer diagnosis with CervixNet. *COGNITIVE SYSTEMS RESEARCH*, *87*.

Stephanie, V., Khalil, I., & Atiquzzaman, M. (2024). DSFL: A Decentralized SplitFed Learning Approach for Healthcare Consumers in the Metaverse. *IEEE TRANSACTIONS ON CONSUMER ELECTRONICS*, *70*(1), 2107–2115.

Subramanian, B., Kim, J., Maray, M., & Paul, A. (2022). Digital Twin Model: A Real-Time Emotion Recognition System for Personalized Healthcare. *IEEE ACCESS*, *10*, 81155–81165.

Suchetha, M., Preethi, S., Veluvolu, K. C., & Raman, R. (2024). An insight in the future of healthcare: integrating digital twin for personalized medicine. *HEALTH AND TECHNOLOGY*, *14*(4), 649–661.

Sun, T., He, X., & Li, Z. (2023). Digital twin in healthcare: Recent updates and challenges. *Digital Health*, *9*. SAGE Publications Inc.

Tao, K., Lei, J. C., & Huang, J. (2024). Physical Integrated Digital twin-based Interaction Mechanism of Artificial Intelligence Rehabilitation Robots Combining Visual Cognition and Motion Control. *WIRELESS PERSONAL COMMUNICATIONS*.

Upreti, D., Yang, E., Kim, H., & Seo, C. (2024). A Comprehensive Survey on Federated Learning in the Healthcare Area: Concept and Applications. *CMES - Computer Modeling in Engineering and Sciences*, *140*(3), 2239–2274. Tech Science Press.

Vallée, A. (2023). Digital twin for healthcare systems. *FRONTIERS IN DIGITAL HEALTH*,

5.
Venkatesh, K. P., Brito, G., & Boulos, M. N. K. (2024). Health Digital Twins in Life Science and Health Care Innovation. *ANNUAL REVIEW OF PHARMACOLOGY AND TOXICOLOGY*, *64*, 159–170.
Vidovszky, A. A., Fisher, C. K., Loukianov, A. D., Smith, A. M., Tramel, E. W., Walsh, J. R., & Ross, J. L. (2024). Increasing acceptance of AI-generated digital twins through clinical trial applications. *CTS-CLINICAL AND TRANSLATIONAL SCIENCE*, *17*(7).
Wang, X., Yu, H., McGee, W., Menassa, C. C., & Kamat, V. R. (2024). Enabling Building Information Model-driven human-robot collaborative construction workflows with closed-loop digital twins. *Computers in Industry*, *161*. Elsevier B.V.
Wu, C., Lorenzo, G., Hormuth, D. A., Lima, E. A. B. F., Slavkova, K. P., DiCarlo, J. C., Virostko, J., et al. (2022). Integrating mechanism-based modeling with biomedical imaging to build practical digital twins for clinical oncology. *Biophysics Reviews*, *3*(2). American Institute of Physics.
Yang, E. (2023). Implications of immersive technologies in healthcare sector and its built environment. *Frontiers in Medical Technology*, *5*. Frontiers Media SA.
Yang, T.-Y., Chien, T.-W., & Lai, F.-J. (2022). Citation analysis of the 100 top-cited articles on the topic of hidradenitis suppurativa since 2013 using Sankey diagrams: bibliometric analysis. *Medicine*, *101*(44), e31144. LWW.
Zackoff, M. W., Rios, M., Davis, D., Boyd, S., Roque, I., Anderson, I., NeCamp, M., et al. (2023). Immersive Virtual Reality Onboarding using a Digital Twin for a New Clinical Space Expansion: A Novel Approach to Large-Scale Training for Health Care Providers. *JOURNAL OF PEDIATRICS*, *252*.
Zhang, J., & Tai, Y. (2022). Secure medical digital twin via human-centric interaction and cyber vulnerability resilience. *Connection Science*, *34*(1), 895–910. Taylor and Francis Ltd.
Zhou, X., Chen, Y., Miao, G., Guo, Y., Zhang, Q., & Bi, J. (2024). Computer-aided robotics for applications in fracture reduction surgery: Advances, challenges, and opportunities. *Iscience*. Elsevier.
Zinchenko, K., & Song, K.-T. (2021). Autonomous Endoscope Robot Positioning Using Instrument Segmentation with Virtual Reality Visualization. *IEEE Access*, *9*, 72614–72623. Institute of Electrical and Electronics Engineers Inc.
Zsidai, B., Hilkert, A. S., Kaarre, J., Narup, E., Senorski, E. H., Grassi, A., Ley, C., et al. (2023). A practical guide to the implementation of AI in orthopaedic research - part 1: opportunities in clinical application and overcoming existing challenges. *JOURNAL OF EXPERIMENTAL ORTHOPAEDICS*, *10*(1).
Zhang, K., Zhou, H. Y., Baptista-Hon, D. T., Gao, Y., Liu, X., Oermann, E., ... & Wu, J. (2024). Concepts and applications of digital twins in healthcare and medicine. *Patterns*, 5(8).

https://doi.org/10.51408/issi2025_076

Document Coverage and Citation-Based Indicators: A Case Study on The Scientific Production of The Federal University of Rio De Janeiro Recovered by Web of Science, Scopus, Dimensions and Lens

Gabriel Alves Vieira[1], Jacqueline Leta[2]

[1]*gabriel.vieira@bioqmed.ufrj.br*, [2]*jleta@bioqmed.ufrj.br*
Federal University of Rio de Janeiro, Institute of Medical Biochemistry Leopoldo de Meis, Zip Code 21.941-590, Rio de Janeiro (Brazil)

Abstract

With the growth of scientific production, quantitative indicators - such as the number of articles published in specialized journals - have assumed an increasingly central role in the evaluation of research institutions, directly influencing the allocation of resources for projects and scholarships. These indicators are directly influenced by the characteristics of the information sources used for their calculation. This study aims to investigate the impact of academic database selection on the calculation of a range of scientific output measures for a single institution: the Federal University of Rio de Janeiro (UFRJ). Four multidisciplinary bibliographic databases were selected for the retrieval of their entire set of UFRJ-related documents: Scopus, Web of Science, Dimensions and Lens. In total, 376,281 documents were retrieved and analyzed using R software. The comparative analyses performed on this corpus include an assessment of UFRJ's scientific production coverage and calculation of citation-based indicators (h, e, g, h^c and i10 indices). The coverage analysis indicates a remarkably high overlap in the corpus retrieved by each source: 28% of the total documents analyzed are covered by all four sources, a percentage that increases to 36% for articles and to 49% for highly cited articles. This suggests that database size is not necessarily a critical factor in selecting an information source for scientific output analysis, especially in contexts where the focus is primarily on journal articles. Furthermore, citation-based indicators exhibited substantial variation both across databases and among the indicators themselves. Notably, a larger number of indexed documents did not necessarily correspond to higher indicator values. These findings indicate that both database choice and citation metrics selection can significantly influence the outcomes of institutional evaluation. It is therefore crucial that managers and professionals engaged in such assessments possess a thorough understanding of the characteristics and limitations of the diverse range of academic databases currently available. This knowledge is essential for selecting appropriate sources and indicators for each situation.

Introduction

In the 1960s, with the growth of global scientific production, objective initiatives towards science evaluation became relevant to Science and Technology (S&T) managers. During this period, the OECD's Frascati manual and other instruments were developed by international bodies to promote the standardization of input and output indicators. In particular, output indicators - which measure the production of S&T documents - have increasingly assumed a central role in defining government policies (Velho, 2001), especially after the emergence of bibliographic databases focused on academic output.

The creation of academic databases - a type of secondary source that indexes metadata from ?? scientific literature (Grogan, 1970) - has significantly boosted scientometric research, which, among other objectives, aims to investigate and quantify the performance and impact of academic research (Mingers & Leydesdorff, 2015; Aria & Cuccurullo, 2017). Scientometric indicators have also been widely adopted in scientific output evaluation processes, as well as in decision-making and policy formulation by S&T managers (Mingers & Leydesdorff, 2015).

Among scientometric indicators, citation-based indicators play a particular role in the assessment of scientific production. These metrics influence not only the ranking of academic journals - often evaluated using citation-based measures (Guerrero-Bote & Moya-Anegón, 2012) - but also the advancement (or ascension) in scientific careers, as funding decisions for research projects and scholarships in many countries frequently involve evaluation processes that incorporate citation-focused indicators (Carlsson, 2009; Schneider, 2009; De Oliveira & Amaral, 2017).

A popular citation-based indicator is the h-index, which is defined as the number h of publications that have each received at least h citations (Hirsch, 2005). For example, an h-index of 25 indicates that the corpus contains 25 publications with at least 25 citations each. The set of documents that contribute to the h-index is named as the h-core, comprising the most highly cited publications within the analyzed corpus. Although originally developed to assess individual researchers, the h-index can be calculated for any collection of documents (Jones et al., 2011), making it a versatile metric for evaluating scientific output at various levels.

Over time, the h-index has inspired the development of several related indicators tailored to specific analytical needs. The *g-index* (Egghe, 2006), for example, is more sensitive to highly cited publications, while the h^c*-index* (Sidiropoulos et al., 2007) gives greater weight to citations received by recently published documents and the *e-index* (Zhang, 2009) differentiates h-cores based on their total citation counts.

Various indicators have also been employed in the construction of university rankings, which are typically elaborated by commercial publishers and publicized as tables that rank higher education institutions based on their performance - an assessment largely driven by quantitative data (Usher & Savino, 2009). Although these rankings are primarily targeted at the general public (such as prospective students seeking a university to attend), they also attract considerable interest within universities themselves, where they may be utilized for auditing, benchmarking, and management purposes (Johnes, 2018).

However, the use of quantitative indicators to evaluate institutional output is far from straightforward, as the choice of metrics and the weight assigned to each can significantly influence ranking outcomes, as noted by Vanz et al. (2018). Moreover, there is evidence that relying on a single database to construct these rankings can introduce bias, owing to variations in coverage across different information sources (Huang et al., 2020). Consequently, bibliographic database selection represents a critical step in the elaboration of academic rankings.

Metrics used in academic evaluation are also directly influenced by the choice of data sources (Gingras, 2016), as databases vary widely in their characteristics, structure, and coverage. For example, different databases employ distinct approaches

to document retrieval and indexing. Bibliographic databases, such as Web of Science (WoS), tend to apply strict selection criteria for the inclusion of new journals into their collections, whereas search engines, like Google Scholar, rely on web crawlers to index vast amounts of academic content available online, aiming for maximum coverage. Additionally, academic databases differ in their thematic scope: while some are multidisciplinary (e.g., Dimensions, Scopus), others specialize in specific fields, such as PubMed for the biomedical sciences or ERIC for education.

Metadata also varies across data sources. This is particularly evident in how academic disciplines are attributed to documents: databases often adopt distinct strategies for this classification, which can generally be divided into two approaches—those that assign disciplines based on the thematic scope of the publication venue, and those that classify documents directly through content analysis (Bornmann, 2018). Another field that frequently differs between sources is document type as each database typically employs its own classification scheme for categorizing the nature of the documents it indexes.

The differences among databases make their selection one of the most crucial steps in the design of any scientometric study aimed at analyzing scientific output. The growing diversity of academic databases, coupled with the need to identify the most appropriate informational source for a given purpose, has given rise to an impressive body of comparative studies examining various secondary sources. A sizable portion of the literature on bibliographic data sources focuses on comparisons between the long-established Web of Science (WoS) and Scopus databases (Archambault et al., 2009; Vieira, Gomes, 2009; Chadegani et al., 2013; Zhu, Liu, 2020). However, the introduction of new academic data platforms, including Dimensions, OpenAlex, and The Lens Scholarly Search, has prompted more recent studies to incorporate these emerging secondary sources into their comparative evaluations (Bornmann et al., 2021; Liang et al., 2021; Delgado-Quiros et al., 2023). Among the topics covered by such studies, the issue of coverage stands out as one of the most analyzed, whether at journal-level (Grindlay et al., 2012; Mongeon, Paul-Hus, 2016; Singh et al., 2021) or, more frequently, document-level (Gusenbauer, 2019; Huang et al., 2020; Martín-Martín et al., 2021; Visser et al., 2021; Gusenbauer, 2022).

The comparative analysis of citation-based indicators across various bibliographic databases serves as a valuable framework for evaluating the relationship between information sources and metrics. This approach facilitates the identification of discrepancies inherent in both the databases and the indicators themselves. Nevertheless, the existing literature on this subject is limited and outdated, frequently focused on specific disciplines and comparing a small number of databases (Franceschet, 2009). Furthermore, to our knowledge, no studies have yet explored these indicators alongside characteristics such as database coverage. Thus, this paper aims to investigate the variations in output retrieved from several databases and the impact of secondary source selection on citation-based indicators. We have opted to conduct a case study that focuses on the scientific output of a single university over its entire publication history which allows us to elucidate the effects of database selection on the assessment of institutional performance across an extended timeframe.

This work analyzes publications from the Federal University of Rio de Janeiro (UFRJ). Founded in 1920, UFRJ stands as Brazil's largest and oldest public university (Oliveira, 2019). The institution offers 176 undergraduate courses and 114 postgraduate programs (PPGs), thereby contributing to the training of professionals and the advancement of research across multiple scientific disciplines. Moreover, it ranks among the top academic institutions in Latin America, being at the 6th position in the 2024 Quacquarelli Symonds university ranking(QS, 2024) and 11th in the 2023 Times Higher Education ranking (THE, 2023). In light with its stature, UFRJ is increasingly focused on enhancing its visibility through strategic investments, including the creation of the Performance Indicator Management Office (GID - https://pr2.ufrj.br/gid), which aims to collect data for university rankings and formulate recommendations for improving the institution's classifications.

Therefore, UFRJ's relevance for Brazilian higher education and scientific development, alongside its extended publication period and increasing focus on factors influencing its standing in academic rankings, justifies its selection as our case of study. Here, we examine variations in UFRJ's scientific output across databases through two main approaches: (i) a comparative analysis of the production retrieved in multiple databases and their coverage; and (ii) an assessment of citation-based indicators calculated for each database.

Methodology

This study was conducted in four main stages: (a) database selection; (b) data collection; (c) data processing; and (d) data analysis. These are presented in the sections below.

Definition of databases

Since the reliability of our results is closely related to document retrieval accuracy, we opted against using academic search engines (e.g., Google Scholar), which tend to exhibit inconsistencies in the results yielded by the same research strategy (Gusenbauer, 2019). Considering the varied scientific output from UFRJ, it seems reasonable to assume that multidisciplinary databases are the most suitable for obtaining a representative sample of the research related to the institution.

We selected four databases: Scopus, Web of Science, Dimensions and Lens. The compatibility of the selected sources with the R bibliometrix package (Aria, Cuccurullo, 2017) was a critical factor in the selection process, as it offers the benefit of automating certain time-consuming steps of the data analysis process. Unfortunately, only these four multidisciplinary databases were supported by the package at that time.

Scopus and Web of Science are the oldest and, selective databases widely utilized in scientometric research (Baas et al., 2020; Birkle et al., 2020), whereas Lens and Dimensions are more recent databases that incorporate third-party sources (Delgado-Quirós, Ortega, 2024) and are less stringent in their indexing criteria. Thus, the selected databases also provide insights into the differences between the two distinct database models.

Data collection

For this stage, all documents indexed by the four sources with at least one author affiliated to UFRJ were retrieved between the last week of January and the first half of February 2023. As the focus was on the institution's scientific output, technological output (e.g., patents) was not retrieved.

The Scopus and Web of Science databases were accessed through the CAPES Periodicals Portal (https://www.periodicos.capes.gov.br). We obtained unrestricted access to the Dimensions interface through its scientometric research support policy (https://www.dimensions.ai/scientometric-research/). Finally, the Lens' academic production retrieval interface (https://www.lens.org/lens/search/scholar/list) required only the creation of a free login to obtain the documents of interest.

All documents were retrieved from the databases' web interfaces. To avoid the inclusion of false positives, the unique identifier 'Affiliation ID' (AF-ID) was used in the Scopus search strategy. Web of Science, on the other hand, has an Affiliation Index that associates variant terms to a canonical institution name. Similarly, the 'Research Organization' field in Dimensions associates all variant terms with a standard institutional name. For Lens, the 'Author Affiliation Name' filter was used with the terms "Federal University of Rio de Janeiro", "UFRJ" and "Federal University of Rio de Janeiro" to select only documents linked to the institution. As the databases restrict the number of publications that can be exported in a single download, we separated the documents into smaller subsets using filters and downloaded them in separate files.

Data processing

Once all UFRJ documents had been collected from the four sources, it was necessary to standardize the data due to the discrepancies observed across various fields in the studied databases. The "convert2df" function of the bibliometrix package was used for that end. This function automatically merges all files obtained for a given database into a single table and standardizes multiple fields.

While bibliometrix tools facilitate semi-automated analysis, further standardization was occasionally required to enhance comparison and visualization. One example was the "Document type" field, which is available in all sources, but features a wide variation in the number of categories used in each database to characterize their documents (Dimensions - 5; Scopus - 15; Lens - 18; and WoS - 27). Thus, using the standard classifications of the sources and comparisons between them would not have been feasible. This problem was solved by reducing the number of document types of all the databases to six common categories: (i) Articles; (ii) Books and book chapters; (iii) Event proceedings; (iv) Preprints; (v) Other; (vi) Unidentified. Table 1 presents the category mapping performed to obtain this standardized classification between the different sources.

Table 1. Merging document categories into a new standardized classification.

Original categories	New categories
'Article', 'Journal article', 'Review', 'Article in press', 'Article; Early access', 'Review; Early access', 'Article; Data paper', 'Data paper'; 'Article; Retracted publication', 'Article; Data paper; Early access', 'Reprint'	Articles
'Proceeding', 'Conference proceedings article', 'Conference proceedings', 'Conference paper', 'Conference review', 'Proceedings paper', 'Meeting abstract', 'Article; Proceedings paper'	Proceedings items
'Book', 'Book chapter', 'Chapter', 'Article; Book chapter', 'Review; Book chapter'	Books and book chapters
NA	Unidentified
'Preprint'	Preprint
'Editorial material', 'Letter', 'Editorial', 'Note', 'Erratum', 'Book review', 'Correction', 'Short survey', 'Report', 'Other', 'Monograph', 'Biographical-item', 'Dataset', 'Abstract report', 'Discussion', 'Clinical trial', 'Dissertation', 'Reference entry', 'News item', 'Correction, Addition', 'Item about an individual', 'Journal issue', 'Bibliography', 'Editorial material; Book chapter', 'Record review', 'News', 'Art exhibit review', 'Chronology', 'Poetry', 'Retraction'	Other

Subsequently, a data cleaning process was conducted to eliminate duplicates in the retrieved dataset, which could potentially lead to an overestimation in the results of subsequent analyses. Documents were grouped as duplicates when: (i) all their fields are identical; (ii) they have a duplicate DOI; or (iii) they present identical information for the title, source, author and publication year fields simultaneously.

Data analysis

Following the standardization and cleaning of the data, the analysis phase began. We adopted a descriptive statistics methodology that primarily leverages totals and percentages to illustrate and summarize various aspects of the corpus retrieved from each database. The *biblioAnalysis* and *summary* functions, both present in bibliometrix, were used to obtain an initial set of statistics, enabling comparisons among different sources. The entire data analysis and visualization processes were performed using the R language v.4.1.2 (R Core Team, 2023) and the Tidyverse metapackage (Wickham et al., 2019).

For the comparative analysis of document distribution between the sources, we used the R package biblioverlap (Vieira & Leta, 2024). This tool processes two or more bibliographic databases, categorizing documents based on the presence or absence of a unique identifier (such as DOI) and detecting document overlap between datasets when: (i) the identifier is identical for two documents; or (ii) the analysis of Title, Author, Year of Publication and Source fields yields a score that surpasses a specified threshold. The package was used to perform a coverage overlap analysis - mapping documents retrieved from distinct databases to pinpoint those appearing in multiple sources at once - on the entire collection of documents and relevant data subsets corresponding to distinct document types. As database classification discrepancies may lead to the pairing of different document types, we used biblioverlap's *get_all_subset_matches* function to retrieve all paired documents against the subsets of interest before analyzing their overlap.

We used only articles for the comparative analysis of citation-based indicators between databases, as these documents are generally the main source of information in citation analysis (Mingers, Leydesdorff, 2015). This variable was used to compute the following indicators: h (Hirsch, 2005); e (Zhang, 2009); g (Egghe, 2006); h^c (Sidiropoulos et al., 2007) and i10. A review by Garner et al. (2018) defines and presents information about the formulae of all these metrics.

Given that citations increase over time and correlate with the availability of citing documents (Tahamtan, 2016), we also aimed to examine the impact of citation windows and the growth of literature size on these indices. First, we split the articles into five groups according to their publication years: one group for those published before 1983 and four additional groups corresponding to each decade from 1983 through 2022. Then, eight metrics of interest were calculated for the set of documents published in each period, namely: (i) number of articles; (ii) total citations received; (iii) average citations received; (iv) h-index; (v) e-index; (vi) g-index; (vii) h^c-index; and (viii) i10-index.

The scripts for data processing and analysis can be found in a public GitHub repository (https://github.com/gavieira/database_coverage_ufrj), which contains thoroughly annotated code that elucidates each step conducted in the process. As Lens allows the redistribution of its data (https://about.lens.org/policies/#acceptableuse), the dataset used in this work can be accessed at https://zenodo.org/records/10500802. The datasets downloaded from the other databases are proprietary and, as such, are not available.

Results

The comparative analyses of UFRJ's scientific output from the selected secondary data sources are organized into three principal sections: (i) an overview of the total number of documents retrieved per database, categorized by publication year and document type; (ii) a document-level coverage analysis of UFRJ's scientific output and the overlap among databases across several data subsets; and (iii) a comparative analysis of citation-based indexes derived from journal articles within each database, examined both collectively and by decade.

Production retrieved and total publications by year and document type

We began our data analysis by examining the total number of UFRJ-affiliated documents retrieved by each datasource. Lens indexed a significantly higher number of documents (113,771) compared to the others. Scopus and Dimensions recovered an intermediate number of items (94,472 and 89,327, respectively), while Web of Science (WoS) returned the smallest set (77,143). Altogether, a total of 374,713 documents were recovered.

The next step was to analyze the annual output of UFRJ across these sources. Figure 1 displays the relative frequency of publications by year and data source. In the chart, each bar represents 100% of the documents in a given year, with the proportion of documents from each source differentiated by color. The value within each colored segment indicates the number of publications indexed by the corresponding source in that year. Notably, overlapping documents (i.e., those indexed by multiple sources) are counted in each segment.

Scopus and Dimensions retrieved the most documents published before 1960, although the overall volume of publications during this early period was relatively small. It is also worth noting that UFRJ's scientific output is covered by all four sources only from 1966 onward, when each database indexed the same number of publications (n = 2). From 1967 to 1970, Dimensions and Lens alternated as the source with the highest number of indexed documents.

Between 1971 and the mid-1980s, documents jointly indexed by Web of Science and Scopus constituted a particularly notable portion of the total. From the mid-1980s to around 2000, the number of documents indexed by each database remained fairly consistent. From 2001 onward, Lens indexed a larger share of UFRJ’s output, except for 2022, the last year of our analysis, when Dimensions and Lens both retrieved more documents than the other sources. Also in 2022, all databases recorded a decline in total publications compared to the previous year.

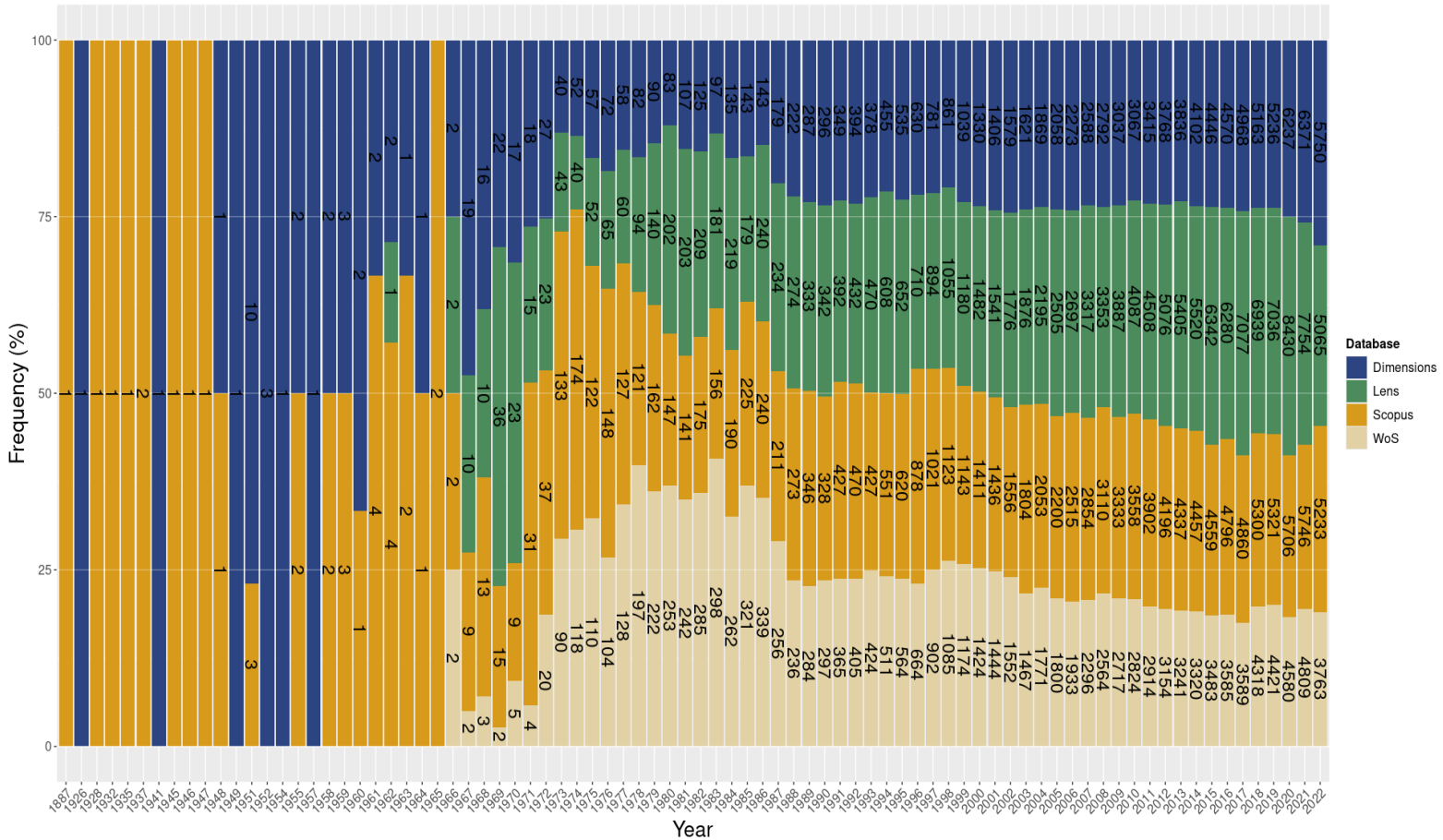


Figure 1. Relative and total frequency of UFRJ documents by year and data source (1887-2022).

As previously mentioned, the distinct document type classifications from each database were unified into six standardized categories, enabling a comparative analysis of their occurrence. The total number of documents per category in each secondary source is presented in Figure 2.

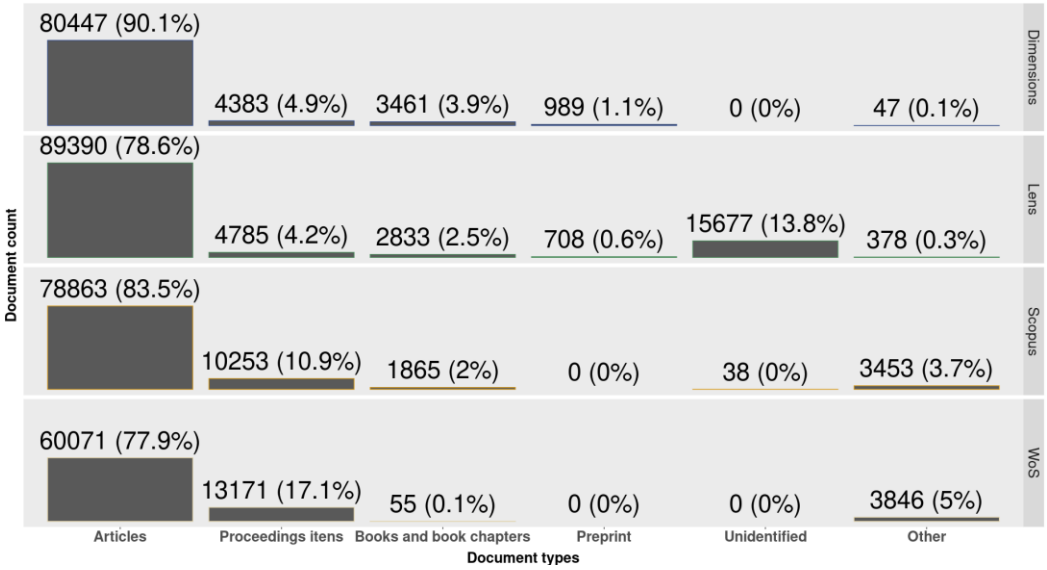


Figure 2. Total UFRJ documents per standardized document category and data source (1887-2022).

Most of the documents consist of articles in all databases, ranging between 77.9% and 90.1% of their respective corpus. Regarding the total number of articles, Lens leads with 89,390 indexed items, surpassing Dimensions 80,447 by nearly 9,000, while Scopus follows closely with 78,863 articles, and Web of Science trails with significantly fewer at 60,071. The scenario is quite different when considering the proceedings items. Web of Science indexes the largest number of documents (13,198), followed by Scopus (10,285), whereas Lens and Dimensions contain substantially fewer at 4,787 and 4,400, respectively.

As for books and chapters, Dimensions leads as the most extensive indexer with 3,461 entries, followed by Lens with 2,833, Scopus with 1,865, and Web of Science, where this document type is nearly non-existent, with only 55 indexed records. Preprints are found only in Dimensions (989) and Lens (708), representing a minor fraction of their documents. Unidentified records are predominantly found in Lens, where they are the second most numerous document type, with 15,677 entries, almost 14% of the documents recovered from the database. Scopus also features a small number of unidentified records (38). Finally, the "Other" category is much more prevalent in Web of Science (3,846) and Scopus (3,453) than in Lens (378) and Dimensions (47).

Coverage analysis by document type

The *biblioverlap* package was employed to identify the extent of overlap in UFRJ's scientific output across the four datasets. All 374,713 retrieved documents were submitted for analysis. All databases combined would yield 164,366 distinct records, provided overlaps are merged into single entries. Of these, 69,285 documents were found to be exclusive to a single data source, while 95,081 appeared in multiple sources. Among the overlapping documents, 92,314 were matched via DOI, whereas the remaining 2,767 were identified through comparative analysis of other bibliographic fields - specifically, title, publication year, first author's name, and journal title.

The results of this coverage analysis were also used to generate Venn diagrams at three distinct aggregation levels: (i) the complete dataset; (ii) subsets based on document type; and (iii) the subset containing the most cited articles, defined here as those that belong to the h-core of each database. The diagrams obtained are shown in Figure 3 and illustrate the document-level coverage overlap in all databases examined. The analysis includes the full set of retrieved documents (3A), records classified as 'articles' (3B), h-core articles for each source (3C), items categorized as 'conference items' (3D), 'books and chapters' (3E), and 'other' types (3F). Each intersection displays the number of documents it contains, followed by the percentage this represents relative to the total number of distinct records analyzed (164,366). The shading of each intersection reflects the number of documents it contains: darker shades correspond to higher values relative to other intersections, while lighter shades to lower values.

The analysis of the full dataset (Fig. 3A) reveals that over a quarter of the documents (45,242) are present across all four databases, whereas those shared between two or three databases are significantly fewer, not exceeding 5% of the total distinct documents. The only exceptions are the document sets found concurrently in Lens, Scopus, and Dimensions (14,021 - 9%) and those in Lens and Dimensions (13,651 - 8%). Regarding documents that occur exclusively in one database, Lens leads with 31,272 records (19%), surpassing even the combined counts from Scopus (16,110 - 10%) and Web of Science (14,819 - 9%), whereas Dimensions contains the fewest exclusive items at 7,084 (4%).

For articles (Fig. 3B), there is a considerable decrease in the proportion of exclusive documents from Scopus, Web of Science, and, above all, Lens. Also, the fraction of articles that occur simultaneously in all sources is bigger (from 28% when analyzing all documents to 36% when analyzing only articles). For h-core articles (Fig. 3C), the percentage of items shared by all datasets is even higher (49%).

Regarding conference items (Fig. 3D), a substantial proportion (11%) is found across all four databases. The only other intersection with a notable share is the one comprising Dimensions, Lens, and Scopus (12%). Beyond these two cases, the presence of conference items in multiple sources is relatively limited, with no other intersection exceeding 5% of the total in this subset. Exclusivity is also prominent in this category, particularly in Scopus (14%) and Web of Science (34%), which hold the largest shares of conference items not indexed by other databases.

For books and chapters (Fig. 3E), Scopus and Dimensions stand out with relatively high proportions of exclusive content - 17% and 13%, respectively -, followed by

Lens at 9%. Conversely, only a very small portion of these documents (1%) is shared across all four databases. The intersection encompassing Lens, Scopus, and Dimensions accounts for the largest share within this category, representing 33% of the subset.

Finally, the 2,107 documents classified under the 'Other' category are simultaneously indexed by all databases (Fig. 3F) - a figure notably higher than the totals reported in the original classifications provided by Lens and Dimensions.

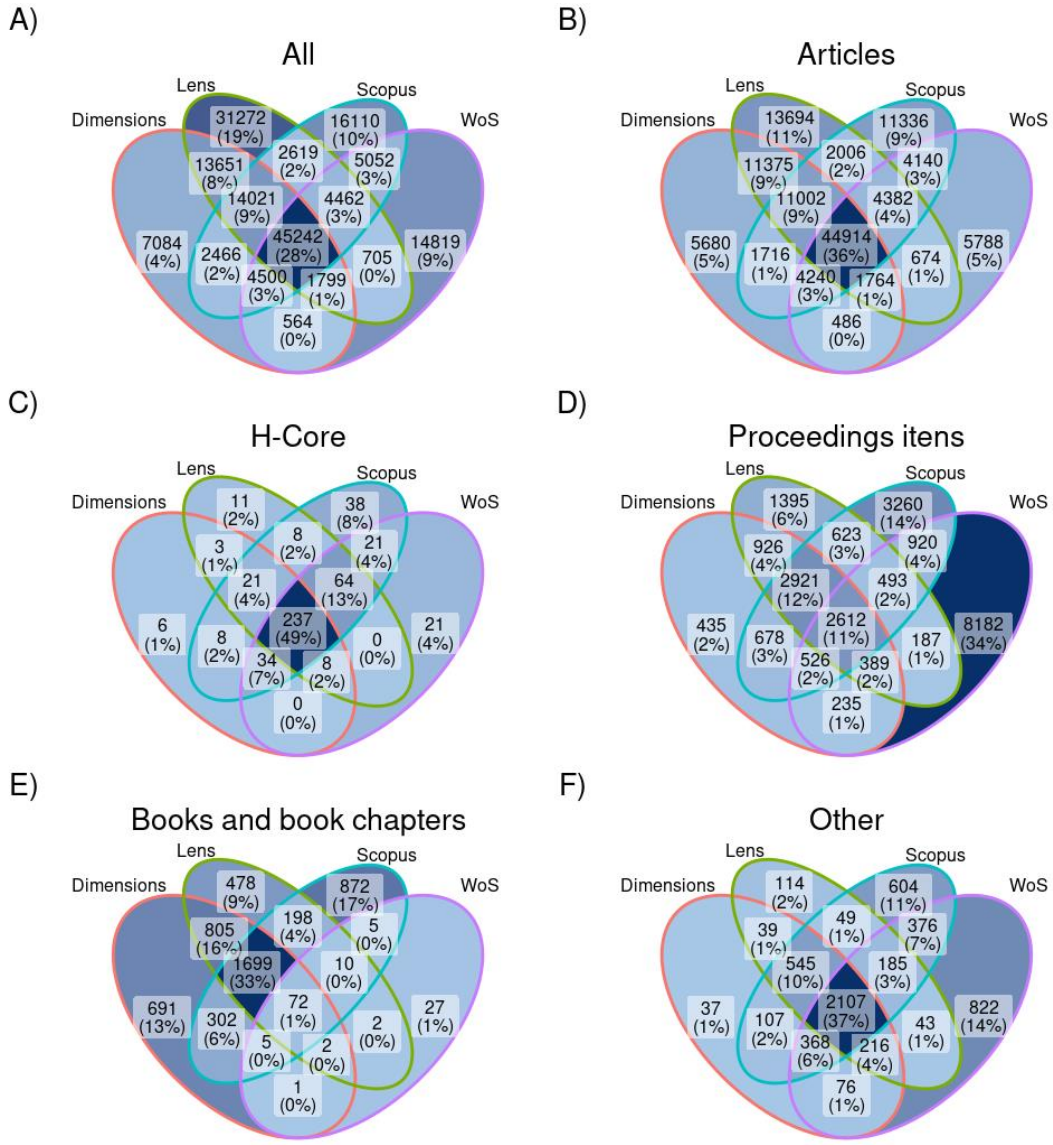


Figure 3. Venn diagrams representing UFRJ's scientific production overlap between the bibliographic databases for multiple data subsets (1887-2022).

Citation-based indicators in each source

In addition to the comparative analysis on the entire corpus of UFRJ publications, five citation-based indices (h, e, g, h^c and i10) were calculated for the set of articles recovered by each database. Table 2 presents these indices along with the total counts of articles and citations.

In general, a higher total article count does not necessarily translate into higher indicator values: if that were the case, Lens would show the highest values and Web of Science the lowest. However, for most indicators, we found that Scopus yields the highest values, followed by Lens, Web of Science, and finally Dimensions. A particularly clear example of this can be seen when comparing Dimensions and Web of Science: although Dimensions has approximately 20,000 more articles and 112,000 more citations than Web of Science, it shows slightly lower values across all indices, except for the i10 index.

Table 2. Total number of items, citations received and visibility indicators calculated based on UFRJ articles retrieved from each data source (1887-2022).

Source	No. of articles	Total citations	Index h	Index e	Index g	Index h^c	i10 Index
Lens	89.390	1.539.910	307	322,18	500	159	33.310
Dimensions	80.447	1.376.751	279	278,05	442	145	32.240
Scopus	78.863	1.642.724	316	353,89	533	164	35.984
WoS	60.071	1.264.978	281	304,86	466	147	28.636

Citation-based indexes and other metrics of interest were also computed by decade (Figure 4). Most of the indices follow a specific trend, regardless of the source: the set of oldest articles (pre-1983) displays the lowest indices, which increase subtly in the period between 1983 and 1992. These indices are significantly higher for articles from the next decade (1993-2002) and continue to grow at the same rate in the period between 2003 and 2012. Then it falls slightly for articles published between 2013 and 2022. Some indicators, such as the i10 index, have diverged from this trend in the last decade analyzed by maintaining their value, while the h^c index has exhibited increments during the same period.

Though all databases follow this pattern, there were still differences in their results. For instance, Scopus performs slightly better than the other databases in most metrics, especially for the last two decades analysed. The exceptions to that are the total number of indexed documents, which are higher in Lens and Dimensions, and mean citations, where WoS outperforms all the other databases.

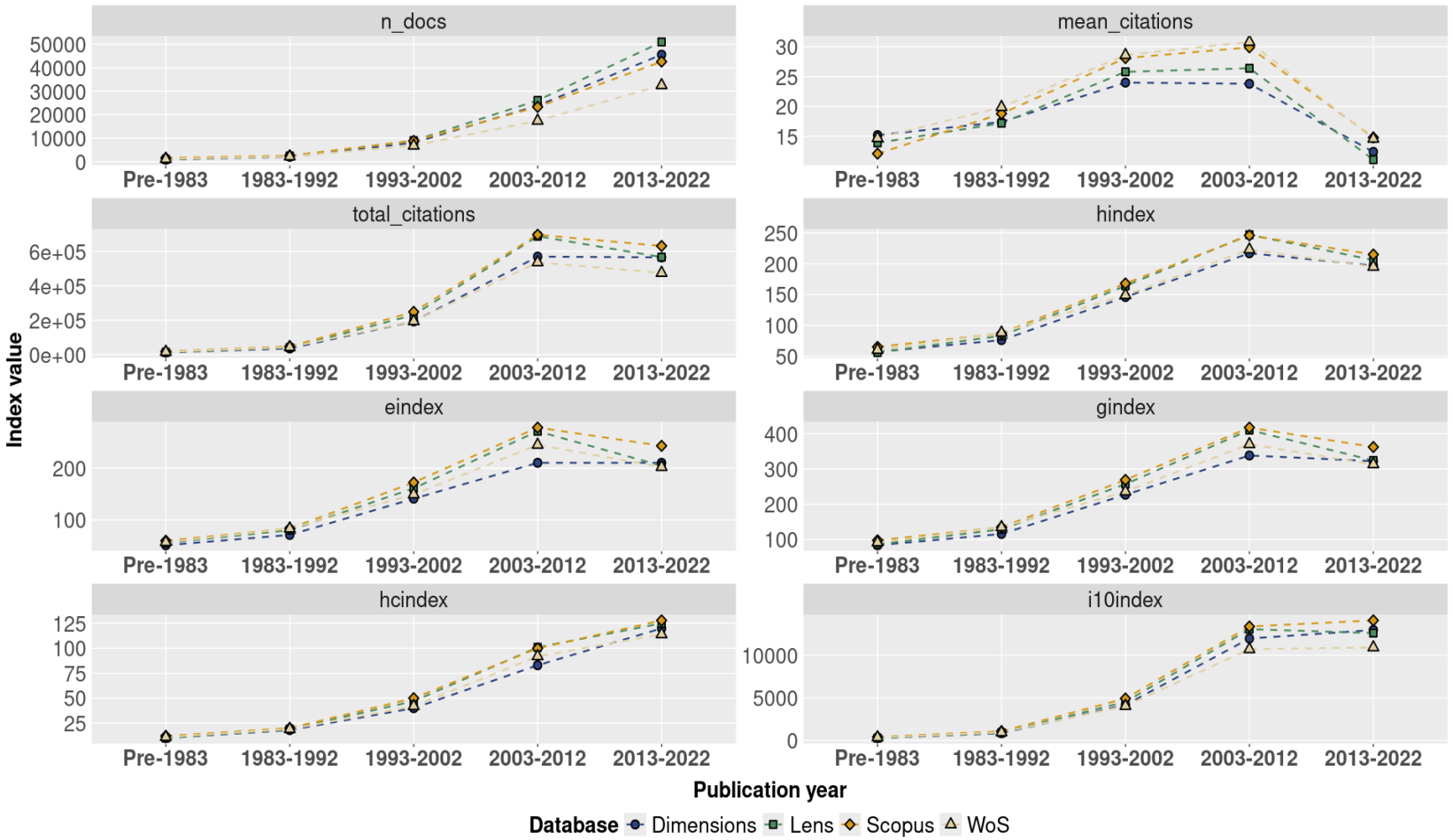


Figure 4. Total number of items, total and average citations received, and visibility indicators calculated for the entire output of UFRJ articles in each data source, grouped by publication year windows (1887-2022).

Discussion

The **total number of documents** retrieved generally correlates with the total indexed documents in each source (Gusenbauer, 2022), indicating that a higher volume of documents increases the likelihood of retrieving a substantial number of relevant documents. However, Dimensions is an exception to this rule. Previous research shows that Dimensions offers significantly broader coverage of publications (Visser et al., 2021) and journals (Singh et al., 2021) compared to Scopus and Web of Science. Notwithstanding, Dimensions retrieved 5,000 less UFRJ-associated documents than Scopus.

This unexpected result may be linked to the substantial proportion of Dimensions documents that are either unaffiliated with any country or institution or lack complete affiliation data (Guerrero-Bote et al., 2021). As this work depends on the quality of the affiliation for data retrieval, it makes sense that Dimensions returned fewer results than Scopus, even though it indexes more documents.

Regarding the distribution of **documents by year**, Scopus and Dimensions retrieve the highest number of records published before 1960. While these documents represent only a small fraction of each database's total corpus, they may be valuable for historiographic studies (Thakuria et al., 2024; Ullah et al., 2023) and related fields. A particularly noteworthy case is a publication from 1887 titled *"The Genesis of the Diamond"*, retrieved from Scopus (Derby, 1887). This stands out because UFRJ was founded in 1920, so any publication predating that year could suggest a metadata error. However, upon examining the document, it was found to be a letter published in *Science*, linked to the National Museum - an institution established in 1818 and later incorporated into UFRJ. Thus, the indexing by Scopus is valid, demonstrating a level of curation quality in this database.

The Lens database indexes a larger proportion of documents from the 21st century compared to the other three sources. This percentage has grown steadily in more recent years until it dropped sharply in 2022, when both Dimensions and Scopus retrieved more documents than Lens. This likely reflects the timing of data collection - early 2023 - when none of the databases had fully indexed the previous year's publications. This is supported by the fact that all databases showed a decrease in the number of documents for 2022 compared to 2021, with the decline being especially marked for Lens. Additionally, Lens was the only database to show a drop in the number of documents published in 2021 relative to 2020. This suggests that Lens may have a slower indexing process, which could be a critical factor for assessments focused on recent literature. However, updated data and more detailed analyses would be necessary to confirm this.

It is also worth noting that, despite its known issue of incomplete affiliation data, Dimensions has consistently indexed more UFRJ-affiliated documents than Scopus since 2020. Two potential hypotheses could explain this: (i) improvements in Dimensions' indexing practices, especially for recent publications and/or (ii) an increase in the volume of content indexed by Dimensions in comparison to Scopus, resulting in a greater number of documents even if there were no improvements to its indexing methodology. Again, further analysis would be required to investigate these claims.

The majority of **documents types** in all four databases are classified as journal articles. However, there is a gap of more than 9,000 articles between the database with the highest article count, Lens, and the second-highest, Scopus. This difference is even more pronounced when compared to the Web of Science (WoS), which has roughly a third as many articles as Lens. While a larger volume of indexed articles can be an attractive characteristic when selecting a database for institutional evaluation, it should not be the sole criterion. Other factors, such as the disciplinary focus of the evaluation and the relevance and prevalence of specific document types, must also be taken into account.

Proceedings items offer a compelling example of this point. Scopus and WoS significantly outperform the newer sources in terms of total indexed records - a pattern that contrasts sharply with what is observed for journal articles. This is especially relevant for evaluating disciplines where research dissemination is more dynamic, such as computer science and related subfields like human-computer interaction, where proceedings are a primary channel for communicating new findings (Freyne et al., 2010; Meho & Rogers, 2008). In such fields, a high count of journal articles may not compensate for a poor representation of proceedings.

In contrast, Dimensions and Lens index more documents classified as books or book chapters. WoS, on the other hand, includes very few of these document types. This limitation may impact its effectiveness in evaluating disciplines where books remain a key vehicle for scholarly communication, which is generally the case for the social sciences and humanities (Kousha & Thelwall, 2015; Bornmann et al., 2016; Toledo, 2020).

As for preprints, it is worth highlighting their growing importance as a means of accelerating the dissemination of research results. While not peer-reviewed, preprints have been especially valuable in contexts that require rapid knowledge sharing, such as during the COVID-19 pandemic (Fraser et al., 2021). However, only Dimensions and Lens include preprints, and even then, they account for only a small fraction of each database's contents.

Unidentified documents were detected in both Lens and Scopus, though their number is negligible in the latter. In contrast, Lens contains a considerable proportion of documents (13.8%) that lack an assigned document type, indicating a potential shortcoming in its classification system..

Documents classified as "other" are abundant in WoS and Scopus, but scarce in Lens and, especially, Dimensions. Differences in how databases classify their content may explain this variation, since this category typically aggregates a large number of widely diverse document types that do not have a direct match in all the sources. Likely, some documents labeled as "Other" in WoS and Scopus are assigned to other categories (e.g., "Articles" or "Conference Items") in Dimensions and Lens. A clearer understanding of these discrepancies, however, requires a closer examination through the subsequent coverage analysis.

Numerous studies have highlighted the considerable variation in coverage overlap of scientific output across different bibliographic databases (Gusenbauer, 2019; Huang et al., 2020; Martín-Martín et al., 2021; Visser et al., 2021). Given this methodology's popularity and usefulness, we have employed a **coverage overlap**

analysis to better characterize UFRJ's scientific output in the four databases analyzed.

As expected, most document matches were made via DOI - a result consistent with previous studies, such as Visser et al. (2021), where 80% or more of the matched documents across sources were linked through DOI-based filters. The remaining matches in those studies were based on combinations of bibliographic metadata like first author's surname and year of publication. In our case, more than two thousand documents without a DOI were still successfully matched using other bibliographic fields. This underscores one of the key advantages of the biblioverlap package, which prioritizes matching via a unique identifier (like DOI) but falls back on a scoring mechanism based on multiple metadata fields when such identifiers are absent (Vieira & Leta, 2024). This approach minimizes matching data loss and is particularly valuable in a multi-disciplinary analysis that includes diverse document types, since DOI assignment practices can vary widely across fields (Gorraiz et al., 2016).

Despite being the database with the largest dataset, Lens presented an unexpectedly high number of unique documents (31,272). Upon manual inspection, we have found that 15,492 of these belonged to entries lacking document type classification. It's worth emphasizing two points here: (i) these documents were generally not associated with DOIs, and (ii) to improve computational efficiency, the biblioverlap algorithm assumes that if a document has a unique identifier (such as a DOI), it will be present in all datasets being analyzed. Consequently, some of the documents deemed unique to Lens may have DOIs in other databases, DOIs that Lens failed to capture, potentially inflating the count of supposedly exclusive documents.

The observed decrease in the number of exclusive documents for the "article" type is largely attributable to the exclusion of document types containing high proportions of unique entries, namely the unidentified documents in Lens and the conference items in Scopus and WoS. In fact, around one-third of journal articles are retrievable from any of the databases, and the number of truly exclusive articles is relatively low. This indicates that, for evaluations centered on journal articles, database coverage alone may not be a distinguishing factor. Other aspects - such as metadata quality and available bibliographic fields - may be more relevant when selecting a source.

We also examined highly cited documents in each database's h-core. Notably, half of these (237) were retrieved by all four sources. This aligns with Visser et al. (2021), who showed that more highly cited documents tend to appear across multiple databases. Since citations are influenced by journal prestige (Martin & Irvine, 1983; Bornmann et al., 2012), and highly cited journals are often prioritized for indexing (Garfield, 1999), it's expected that these publications will appear in multiple sources.

Proceeding items show a markedly different pattern: a low overlap across sources and a high number of documents exclusive to one database. WoS retrieves the largest share of these documents (around 57%), and combining WoS and Scopus increases this to approximately 89% of all UFRJ-affiliated conference items. This supports the use of both databases in evaluations of disciplines where proceedings are a key publication venue.

About one-third of books and book chapters were found simultaneously in Scopus, Dimensions, and Lens - a direct consequence of the near absence of this document type in WoS. The limited increase in the total number of WoS records after accounting for matches with other document types supports the conclusion that this is a real coverage gap rather than a classification issue. Thus, WoS may not be suitable for evaluation processes focused on areas where books are particularly relevant for scientific communication.

The “Other” document type presents another intriguing case. In Dimensions and Lens, the number of documents matched to the entries categorized as “Other” by the remaining databases was far greater than the number of items those sources originally classified as such. Manual inspection revealed that the majority of these documents were classified as "Articles" in Dimensions (97.8%) and Lens (88.2%), suggesting classification errors. While further investigation would be needed to determine definitively which databases are misclassifying documents, Dimensions and Lens, being relatively recent and drawing heavily from open data aggregators like PubMed and Crossref (Herzog et al., 2020; Cambia, 2024a), are more likely to be the sources of these inconsistencies. Metadata quality from Scopus and WoS is generally regarded as more reliable (Guerrero-Bote et al., 2021; Delgado-Quirós & Ortega, 2024).

Such classification issues have serious implications for scientometric analyses. As previously discussed, the impact and relevance of document types vary greatly by discipline. Misclassifications can skew evaluations or lead to erroneous conclusions. Therefore, while total document count is an important metric when choosing a data source, harder-to-measure qualities - like metadata accuracy and classification reliability - are just as critical, if not more so.

The h-index combines publication and citation counts into a single metric (Hirsch, 2005), favoring documents with higher citation volumes. This attribute is shared by its derivatives, such as the e-, g-, and hc-indexes. As a result, smaller databases may yield higher values for **citation-based indices** values than larger ones, depending on how well they capture highly cited publications.

Our findings reflect this pattern. WoS outperformed Dimensions, and Scopus outperformed Lens, despite the latter two having broader overall coverage. We suggest two main aspects that could explain this outcome: (i) the more performant databases may include highly cited documents absent from other sources; or (ii) they may have more efficient citation-linking mechanisms. The first seems less likely, as our overlap analysis showed that approximately half of the h-core documents are shared across all platforms.

The second aspect is more plausible. Issues with metadata precision appear to hinder accurate citation tracking in the newer sources. For example, Lens has a substantial number of uncategorized documents, suggesting a lack of granularity in its curation process. Similarly, despite its large document base, Dimensions retrieved fewer publications, likely due to deficiencies in the 'affiliation' field. Visser et al. (2021) corroborates this view by reporting that, while highly cited articles tend to be present in all major databases, WoS and Scopus demonstrate superior citation-linking capabilities. Our results are consistent with these findings.

An exception to this pattern is the i10-index, which counts the number of publications with at least ten citations. Unlike other indices, it is not increasingly difficult to raise its value over time. Because it uses a fixed, relatively low citation threshold, it is also less sensitive to errors in citation linking. Notably, this was the only index where Dimensions outperformed WoS. This underscores the importance of carefully selecting citation metrics, as different indices may produce varying outcomes depending on the data source and characteristics of the dataset.

When analyzing UFRJ's scientific output by decade, we found that citation-based indices are shaped by both the volume and age of publications. Early periods had few indexed articles and lower index values, even though these documents had more time to accumulate citations. Later decades featured both an increase in publication volume and more extensive citation windows, which corresponded to higher index values. In the most recent decade, although publication volume kept growing, the indices plateaued or declined - likely a result of limited time for newer publications to accrue citations.

Interestingly, the i10 and h^c indexes did not decline in the most recent decade. The i10-index's resilience likely reflects its modest citation threshold, though Lens saw a drop that may be linked to slower indexing of recent publications. The hc-index, which gives greater weight to recent citations, actually increased, as expected.

Together, these results demonstrate that citation-based index values are not only influenced by the selected database but are also highly dependent on the metric chosen. Scopus consistently delivered higher index values, likely due to its balanced combination of broad coverage and efficient citation linking. WoS, while similarly strong in citation linking, indexes a more selective subset of publications. This results in higher average citation values but not necessarily higher index values. By contrast, Dimensions and Lens reported lower mean citation values, pointing to either less effective citation tracking, the inclusion of more poorly cited documents, or both.

Conclusion

Although this study offers important insights into how crucial database selection may be to institutional research evaluation, it has several limitations. First, it is based on a single case study and considers only two variables: document counts and citation counts. Furthermore, it does not split the production by discipline and evaluates only four bibliographic databases.

The analyses were conducted using the complete scientific output of one university - UFRJ. We make no claims that these results are generalizable to other institutions, and we recognize that similar analyses may yield different results elsewhere. For transparency and reproducibility, the datasets used (where legally permissible) and the analysis code have been made publicly available on Zenodo and GitHub. We hope this facilitates the application of our analytical framework to other institutions, encouraging replication, validation, or expansion of our findings while addressing different dimensions of institutional evaluation. In fact, we are currently conducting a follow-up study using a similar methodology to examine both high- and low-ranked institutions, aiming to determine whether characteristics such as publication overlap correlate with institutional reputation.

An additional limitation lies in the narrow focus on two variables, document and citation counts, which excludes other relevant dimensions of research output. These include collaboration patterns (both national and international) and adherence to open access publishing models, both of which are available in the databases analyzed. Future studies would benefit from incorporating these additional dimensions alongside citation-based indices. Such multifaceted analyses could support the development of custom indicators or institutional rankings, as seen in the work of Huang et al. (2020).

Another constraint is the absence of field-level classification in our analysis. Categorizing scientific output by discipline is a valuable addition to any bibliometric study, allowing for finer-grained comparisons, particularly when interpreting citation-based indicators. Including such classification in future work - whether by mapping categories across databases (Singh et al., 2021) or through publication-level classification based on content (Rivest et al., 2021; Pech et al., 2022) - would enhance analytical depth.

Disciplinary classification would also allow for normalization of citation-based indicators by field (Waltman & van Eck, 2013) and enable more appropriate comparisons across disciplines. Moreover, it would support the exploration of domain-specific patterns, such as citation half-life (Burton & Kebler, 1960), which varies significantly across research areas. With publications sorted by discipline, it becomes possible to assess how citation window length affects the evaluation of different fields across various sources. Ultimately, this would improve our understanding of how database and indicator choices may affect evaluations not only at the institutional level but also within specific disciplines.

Lastly, the decision to analyze four databases was a methodological choice driven by the scope of the study. We opted to examine the full scholarly output of a large institution from its founding up to the year before data collection. Expanding the analysis to include more sources would have significantly increased the data volume and required much more time for retrieval and processing. However, we intend to include additional databases in future research. OpenAlex - a relatively recent, open-access, multidisciplinary source - stands out as a promising candidate due to its publicly available API and user-friendly interface (Priem et al., 2022).

The primary objective of this study was to examine how the choice of bibliographic database can affect both the set of retrieved publications and the calculation of citation-based indicators, as well as to discuss the broader implications for evaluating an institution's scientific output. To achieve this, we conducted a case study using the complete scholarly production of the Federal University of Rio de Janeiro (UFRJ), as retrieved from four multidisciplinary databases: Web of Science (WoS), Scopus, Dimensions, and Lens.

Our analysis of UFRJ's production across these databases revealed several findings that support the notion that results vary substantially depending on the source, particularly in terms of total document count, document type coverage, and citation-based metrics. One key takeaway is that a larger database does not necessarily guarantee higher retrieval of relevant documents or better citation metrics, as factors such as metadata quality are equally, if not more, important than the number of indexed items.

For instance, although Dimensions is considerably larger than Scopus in terms of overall indexed content (Visser et al., 2021), it retrieved fewer documents in this case study. This is likely due to limitations in the quality of metadata, especially within the 'affiliation' field (Guerrero-Bote et al., 2021), which was used as the main retrieval criterion. Similarly, Lens displayed signs of metadata quality issues, such as a high proportion of records lacking classification by document type. When analyzing citation-based indices, we found that databases with higher retrieval counts did not necessarily yield higher index values: Lens, for example, was outperformed by Scopus, while Dimensions was outperformed by WoS. These findings likely reflect differences in the efficiency of citation link identification across databases.

In summary, although newer data sources, like Dimensions and Lens, tend to be more comprehensive in terms of document indexing, they also show signs of lower metadata quality when compared to more established sources like WoS and Scopus. Therefore, selecting a database solely based on its volume of indexed content may be inadvisable. Because scientific output evaluations can be significantly influenced by both metadata quality and coverage of specific document types, neglecting these characteristics may lead to inaccurate assessments. Similarly, selecting appropriate citation-based indices is essential to avoid distortions caused by, for instance, a few highly cited publications.

Our coverage analysis also showed that a significant proportion of documents - especially journal articles and highly cited papers - appear in all four databases. However, there was wide variation in document type coverage, which is relevant given that different disciplines often rely on different formats for scholarly communication. These results further underscore the inadequacy of database selection based on document count alone.

In the context of our dataset, Scopus emerged as the most suitable database in terms of both document retrieval and citation-based indicators. This finding aligns with the methodologies of prominent university rankings such as QS and THE, both of which use Scopus as the underlying data source for evaluating institutional output. Additionally, Scopus included a large number of items classified as "Books or Chapters," achieving comparable coverage of this document type to that of Dimensions and Lens. Considering the lower metadata quality observed in the newer sources (Guerrero-Bote et al., 2021; Visser et al., 2021), Scopus appears particularly well-suited for evaluating fields in which books constitute a key channel for scholarly communication.

Nevertheless, this does not imply that other databases should be disregarded. Dimensions, for instance, integrates data on publications, altmetrics, clinical trials, patents, funding, and institutional policies, which are interlinked through citations and other types of connections. According to Herzog et al. (2020), these relationships enable a holistic analysis of the scientific production cycle: from initial research funding to publication, technological application, and influence on policy development. The authors also note that the developers of Dimensions are aware of the limitations associated with this database and are actively working with publishers and other partners to enhance its content quality and coverage. The same can likely be said for Lens. Given that both databases were launched in the late 2010s,

significant improvements in their performance and coverage can be expected over time.

Although WoS indexed fewer documents than Scopus, it demonstrated a more granular classification of document types. It also outperformed Dimensions, despite the latter's larger size, which suggests that WoS is highly efficient in establishing citation links among its records. Moreover, WoS's rich metadata and extensive collection of "Conference items" indicate that it is a valuable resource for evaluating disciplines where such formats are a key channel of scientific communication.

We hope that the findings presented here raise awareness among researchers, evaluators, and policymakers regarding how both database and metric selection can significantly affect institutional assessments. Recognizing these effects is a critical step toward promoting higher standards in research evaluation and ensuring that methodologies are appropriately tailored to the specific characteristics and needs of each evaluation context.

Acknowledgements

We would like to thank CNPq for their continued financial support.

References

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. Journal of the American Society for Information Science and Technology, 60(7), 1320–1326. https://doi.org/10.1002/asi.21062

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies, 1(1), 377–386. https://doi.org/10.1162/qss_a_00019

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. Quantitative Science Studies, 1(1), 363–376. https://doi.org/10.1162/qss_a_00018

Bornmann, L. (2018). Field classification of publications in Dimensions: A first case study testing its reliability and validity. Scientometrics, 117(1), 637–640. https://doi.org/10.1007/s11192-018-2855-y

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. Humanities and Social Sciences Communications, 8(1), Article 1. https://doi.org/10.1057/s41599-021-00903-w

Bornmann, L., Schier, H., Marx, W., & Daniel, H.-D. (2012). What factors determine citation counts of publications in chemistry besides their quality? Journal of Informetrics, 6(1), 11–18. https://doi.org/10.1016/j.joi.2011.08.004

Bornmann, L., Thor, A., Marx, W., & Schier, H. (2016). The application of bibliometrics to research evaluation in the humanities and social sciences: An exploratory study using normalized Google Scholar data for the publications of a research institute. Journal of the Association for Information Science and Technology, 67(11), 2778–2789. https://doi.org/10.1002/asi.23627

Burton, R. E., & Kebler, R. W. (1960). The “half-life” of some scientific and technical literatures. American Documentation, 11(1), 18–22. https://doi.org/10.1002/asi.5090110105

Carlsson, H. (2009). Allocation of Research Funds Using Bibliometric Indicators – Asset and Challenge to Swedish Higher Education Sector. Infotrend, 64(4), 82–88.

Chadegani, A. A., Salehi, H., Yunus, M. M., Farhadi, H., Fooladi, M., Farhadi, M., & Ebrahim, N. A. (2013). A Comparison between Two Main Academic Literature Collections: Web of Science and Scopus Databases. Asian Social Science, 9(5), Article 5. https://doi.org/10.5539/ass.v9n5p18

de Oliveira, T. M., & Amaral, L. (2017). Políticas Públicas em Ciência e Tecnologia no Brasil: Desafios e propostas para utilização de indicadores na avaliação. In Bibliometria e Cientometria no Brasil: Infraestrutura para avaliação da pesquisa científica na Era do Big Data (pp. 157–184). ECA/USP.

Delgado-Quirós, L., Aguillo, I. F., Martín-Martín, A., López-Cózar, E. D., Orduña-Malea, E., & Ortega, J. L. (2023). Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases. Journal of the Association for Information Science and Technology, asi.24839. https://doi.org/10.1002/asi.24839

Delgado-Quirós, L., & Ortega, J. L. (2024). Completeness degree of publication metadata in eight free-access scholarly databases. Quantitative Science Studies, 1–19. https://doi.org/10.1162/qss_a_00286

Derby, O. A. (1887). The Genesis of the Diamond. Science, ns-9(207), 57–58. https://doi.org/10.1126/science.ns-9.207.57

Egghe, L. (2006). Theory and practise of the g-index. Scientometrics, 69(1), 131–152. https://doi.org/10.1007/s11192-006-0144-7

Franceschet, M. (2009). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. Scientometrics, 83(1), 243–258. https://doi.org/10.1007/s11192-009-0021-2

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., Nanni, F., & Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. PLOS Biology, 19(4), e3000959. https://doi.org/10.1371/journal.pbio.3000959

Freyne, J., Coyle, L., Smyth, B., & Cunningham, P. (2010). Relative status of journal and conference publications in computer science. Communications of the ACM, 53(11), 124–132. https://doi.org/10.1145/1839676.1839701

Garfield, E. (1999). Journal impact factor: A brief review. CMAJ, 161(8), 979–980.

Garner, R. M., Hirsch, J. A., Albuquerque, F. C., & Fargen, K. M. (2018). Bibliometric indices: Defining academic productivity and citation rates of researchers, departments and journals. Journal of Neurointerventional Surgery, 10(2), 102–106. https://doi.org/10.1136/neurintsurg-2017-013265

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. Journal of Informetrics, 10(1), 98–109. https://doi.org/10.1016/j.joi.2015.11.008

Grindlay, D. J. C., Brennan, M. L., & Dean, R. S. (2012). Searching the Veterinary Literature: A Comparison of the Coverage of Veterinary Journals by Nine Bibliographic Databases. Journal of Veterinary Medical Education, 39(4), 404–412. https://doi.org/10.3138/jvme.1111.109R

Guerrero-Bote, V. P., Chinchilla-Rodríguez, Z., Mendoza, A., & de Moya-Anegón, F. (2021). Comparative Analysis of the Bibliographic Data Sources Dimensions and

Scopus: An Approach at the Country and Institutional Levels. Frontiers in Research Metrics and Analytics, 5. https://www.frontiersin.org/article/10.3389/frma.2020.593494
Guerrero-Bote, V. P., & Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. Journal of Informetrics, 6(4), 674–688. https://doi.org/10.1016/j.joi.2012.07.001
Gusenbauer, M. (2019). Google Scholar to overshadow them all? Comparing the sizes of 12 academic search engines and bibliographic databases. Scientometrics, 118(1), 177–214. https://doi.org/10.1007/s11192-018-2958-5
Gusenbauer, M. (2022). Search where you will find most: Comparing the disciplinary coverage of 56 bibliographic databases. Scientometrics, 127(5), 2683–2745. https://doi.org/10.1007/s11192-022-04289-7
Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. Quantitative Science Studies, 1(1), 387–395. https://doi.org/10.1162/qss_a_00020
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102
Huang, C.-K. (Karl), Neylon, C., Brookes-Kenworthy, C., Hosking, R., Montgomery, L., Wilson, K., & Ozaygen, A. (2020). Comparison of bibliographic data sources: Implications for the robustness of university rankings. Quantitative Science Studies, 1(2), 445–478. https://doi.org/10.1162/qss_a_00031
Johnes, J. (2018). University rankings: What do they really show? Scientometrics, 115(1), 585–606. https://doi.org/10.1007/s11192-018-2666-1
Jones, T., Huggett, S., & Kamalski, J. (2011). Finding a Way Through the Scientific Literature: Indexes and Measures. World Neurosurgery, 76(1), 36–38. https://doi.org/10.1016/j.wneu.2011.01.015
Kousha, K., & Thelwall, M. (2015). Web indicators for research evaluation. Part 3: Books and non standard outputs. Profesional de La Información, 24(6), Article 6. https://doi.org/10.3145/epi.2015.nov.04
Liang, Z., Mao, J., Lu, K., & Li, G. (2021). Finding citations for PubMed: A large-scale comparison between five freely available bibliographic data sources. Scientometrics, 126(12), 9519–9542. https://doi.org/10.1007/s11192-021-04191-8
Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. Research Policy, 12(2), 61–90. https://doi.org/10.1016/0048-7333(83)90005-7
Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. Scientometrics, 126(1), 871–906. https://doi.org/10.1007/s11192-020-03690-4
Meho, L. I., & Rogers, Y. (2008). Citation counting, citation ranking, and h -index of human-computer interaction researchers: A comparison of Scopus and Web of Science. Journal of the American Society for Information Science and Technology, 59(11), 1711–1726. https://doi.org/10.1002/asi.20874
Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. European Journal of Operational Research, 246(1), 1–19. https://doi.org/10.1016/j.ejor.2015.04.002
Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. Scientometrics, 106(1), 213–228. https://doi.org/10.1007/s11192-015-1765-5

Pech, G., Delgado, C., & Sorella, S. P. (2022). Classifying papers into subfields using Abstracts, Titles, Keywords and KeyWords Plus through pattern detection and optimization procedures: An application in Physics. Journal of the Association for Information Science and Technology, 73(11), 1513–1528. https://doi.org/10.1002/asi.24655

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts (No. arXiv:2205.01833). arXiv. https://doi.org/10.48550/arXiv.2205.01833

QS. (2024, July 15). South America QS World University Rankings 2024. Top Universities. https://www.topuniversities.com/latin-america-south-america-rankings

Rivest, M., Vignola-Gagné, E., & Archambault, É. (2021). Article-level classification of scientific publications: A comparison of deep learning, direct citation and bibliographic coupling. PLOS ONE, 16(5), e0251493. https://doi.org/10.1371/journal.pone.0251493

Schneider, J. W. (2009). An Outline of the Bibliometric Indicator Used for Performance-Based Funding of Research Institutions in Norway. European Political Science, 8(3), 364–378. https://doi.org/10.1057/eps.2009.19

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. Scientometrics, 72(2), 253–280. https://doi.org/10.1007/s11192-007-1722-z

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. Scientometrics, 126(6), 5113–5142. https://doi.org/10.1007/s11192-021-03948-5

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. Scientometrics, 107(3), 1195–1225. https://doi.org/10.1007/s11192-016-1889-2

Thakuria, A., Chakraborty, I., & Deka, D. (2024). A bibliometric review on serendipity literature available in Web of Science database using HistCite and Biblioshiny. Information Discovery and Delivery, 52(2), 227–242. https://doi.org/10.1108/IDD-01-2023-0001

THE. (2023, June 12). Latin America University Rankings 2023. Times Higher Education (THE). https://www.timeshighereducation.com/world-university-rankings/2023/latin-america-university-rankings

Toledo, E. G. (2020). Why Books are Important in the Scholarly Communication System in Social Sciences and Humanities. Scholarly Assessment Reports, 2(1), 6. https://doi.org/10.29024/sar.14

Ullah, F., Shen, L., & Shah, S. H. H. (2023). Value co-creation in business-to-business context: A bibliometric analysis using HistCite and VOS viewer. Frontiers in Psychology, 13. https://doi.org/10.3389/fpsyg.2022.1027775

Usher, A., & Savino, M. (2009). A Global Survey of University Ranking and League Tables. In The Routledge International Handbook of Higher Education. Routledge.

Vanz, S. A. de S., Dominique, A. P., Lascurain Sánchez, M. L., & Sanz Casado, E. (2018). Rankings universitários internacionais e o desafio para as universidades brasileiras. Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência Da Informação. Florianópolis, SC. Florianópolis. Vol. 23, n. 53 (Set./Dez. 2018), p. 39-51.

Velho, L. M. S. (2001). Estratégias para um sistema de indicadores de C&T no Brasil. 13.

Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. Scientometrics, 81(2), 587–600. https://doi.org/10.1007/s11192-009-2178-0

Vieira, G. A., & Leta, J. (2024). biblioverlap: An R package for document matching across bibliographic datasets. Scientometrics, 129(7), 4513–4527. https://doi.org/10.1007/s11192-024-05065-5

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies, 2(1), 20–41. https://doi.org/10.1162/qss_a_00112

Waltman, L., & van Eck, N. J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. Journal of Informetrics, 7(4), 833–849. https://doi.org/10.1016/j.joi.2013.08.002

Zhang, C.-T. (2009). The e-Index, Complementing the h-Index for Excess Citations. PLOS ONE, 4(5), e5429. https://doi.org/10.1371/journal.pone.0005429

Zhu, J., & Liu, W. (2020). A tale of two databases: The use of Web of Science and Scopus in academic papers. Scientometrics, 123(1), 321–335. https://doi.org/10.1007/s11192-020-03387-8

https://doi.org/10.51408/issi2025_007

Enhancing Research Idea Generation through Combinatorial Innovation and Multi-Agent Iterative Search Strategies

Shuai Chen[1], Chengzhi Zhang[2]

[1]*shuaichen@njust.edu.cn*, [2]*zhangcz@njust.edu.cn*
Department of Information Management, Nanjing University of Science and Technology, Nanjing, 210094 (China)

Abstract

Scientific progress relies on the continuous emergence of innovative discoveries. However, the exponential growth in scientific literature has increased the cost of information filtering, making it significantly more challenging for scientists to identify innovative research directions. Although artificial intelligence (AI) methods have shown potential in tasks such as research idea generation and hypothesis formulation, the ideas they produce are often repetitive and simplistic. Combinatorial innovation theory posits that new entities arise from the recombination of existing elements, offering a novel approach to addressing these challenges.

This study draws on combinatorial innovation theory and the Delphi method to introduce a multi-agent iterative planning and search strategy into the research idea generation process, aiming to enhance the diversity and novelty of generated ideas. The strategy integrates iterative knowledge search with a large language model (LLM)-based multi-agent system to iteratively generate, evaluate, and refine research ideas. Experiments conducted using data from the field of natural language processing demonstrate that the multi-agent iterative planning and search strategy outperforms state-of-the-art methods in terms of diversity and novelty, showcasing its potential to generate high-quality research ideas.This study not only validates the effectiveness of the multi-agent iterative search strategy but also provides a theoretical explanation, grounded in combinatorial innovation theory and methodologies, for its ability to improve research idea generation performance. It offers new perspectives for future work in this domain.

Introduction

Over the past few decades, the volume of scientific literature has experienced exponential growth, reflecting the vigorous expansion of research activities and the continuous advancement of science and technology. However, the sheer magnitude of scientific publications has imposed significant temporal and cognitive burdens on scientists as they endeavor to filter and assimilate relevant information. Concurrently, this phenomenon has exacerbated the issue of redundancy in scientific research(Larivière et al., 2008), leading to substantial inefficiencies in the allocation of research resources. These compounding factors have collectively contributed to the escalating challenges scientists face in pursuing innovative research endeavors.

Recently, large language models (LLMs) have demonstrated remarkable performance across a variety of challenging tasks, including mathematical proof (Yang et al., 2024), information retrieval(Ajith et al., 2024), and solving specific research problems through code generation(Lu et al., 2024; Schmidgall et al., 2025; Yuan et al., 2025) . These models have even shown the potential to generate innovative research ideas(Baek et al., 2024; X. Gu & Krenn, 2024; Kumar et al., 2024; Lu et al., 2024; Si et al., 2024). However, despite evidence suggesting that LLMs can produce novel research concepts, their outputs often exhibit a high degree

of redundancy(Si et al., 2024). This issue necessitates additional effort from researchers to filter and deduplicate generated content, thereby hindering their broader application in academic research. Several studies have attempted to address this challenge through various approaches. For instance, (Baek et al., 2024) employed knowledge graph construction, (Si et al., 2024) utilized keyword-based searches for specific knowledge, and (Hu et al., 2024) adopted iterative knowledge search strategies. Nevertheless, these methods remain limited in critical ways. On the one hand, they often focus narrowly on knowledge within a single domain, failing to adequately integrate insights from multiple related fields. This significantly constrains the breadth of knowledge sources and the diversity of problem-solving perspectives. On the other hand, these studies have not sufficiently addressed the potential biases introduced by relying on a single large language model.

In light of these considerations, this study introduces combinatorial innovation theory and the multi-agent iterative planning and search strategy to the task of research idea generation. This strategy leverages knowledge planning and search mechanisms to integrate multi-domain knowledge, supported by a large language model-based multi-agent system. It simulates the expert survey method (Delphi method)(Linstone & Turoff, 1975) commonly employed in innovation practices, iteratively generating, evaluating, and refining research ideas. Specifically, the large language model is assigned the role of an expert with a specific disciplinary background to simulate the Delphi method discussion process in real-world research scenarios. Experiments conducted on a dataset of academic papers in the field of natural language processing demonstrate that the proposed method outperforms baseline approaches across key metrics, including diversity, novelty, and quality scores. Furthermore, the study provides an explanation for the enhanced performance of the strategy in generating research ideas, drawing on combinatorial innovation theory and methodological applications. This offers novel insights and perspectives for future research on idea generation. The study addresses the following two research questions:

RQ1: Can the multi-agent iterative planning and search strategy enhance the diversity and novelty of research ideas generated by large language models?

RQ2: Can combinatorial innovation theory and methodological approaches guide the task of generating research ideas using large language models?

The contributions of this study are threefold:

First, this paper proposes a multi-agent iterative planning and search strategy, which is applied to the task of generating research ideas using large language models. The strategy is evaluated through role-playing simulations with real-world data, and the final outputs are assessed objectively.

Second, the study conducts comprehensive experiments to evaluate the multi-agent iterative planning and search strategy. These experiments include comparisons with baseline methods, assessments of different team configurations, variations in the number of iterations, and ablation studies of individual modules. The results demonstrate that the proposed strategy significantly enhances the quality of research idea generation, outperforming existing baseline methods.

Third, this paper provides a theoretical explanation for the improved performance of the multi-agent iterative planning strategy in generating research ideas, drawing on combinatorial innovation theory and methodological applications. This not only offers new insights into the mechanisms underlying the strategy's success but also provides a novel tool and perspective for future research on idea generation using large language models.

Related work

This section reviews related work from three perspectives: (1) Generating Research Ideas Using Large Language Models; (2) Prompt Engineering for Logical Reasoning in Large Language Models; (3) Combinatorial Innovation Theory and Methodological Approaches.

Generating Research Ideas Using Large Language Models

In recent years, a growing body of research has demonstrated that large language models (LLMs) possess the capability to generate novel and innovative scientific research ideas, a phenomenon that has garnered significant attention from scholars. Among these studies, some researchers have adopted approaches such as retrieving relevant papers based on research topics(Lu et al., 2024) or directly utilizing the references of target papers(Guo et al., 2024), embedding these materials into the contextual prompts of LLMs to stimulate the generation of related research ideas. Others have first retrieved relevant papers as a knowledge base and enhanced idea generation by retrieving related knowledge during the process(Si et al., 2024) , a method known as Retrieval-Augmented Generation (RAG) (Lewis et al., 2020). Additionally, some scholars have constructed scientific knowledge graphs, employing co-occurrence entity search techniques to integrate retrieved entities into LLM prompts, thereby generating unique and novel research ideas(Baek et al., 2024; X. Gu & Krenn, 2024). IdeaSynth(Pu et al., 2024) introduced human expertise into the research idea generation process, demonstrating that human-AI collaboration outperforms single LLM baselines. VIRSCI(Su et al., 2024) further incorporated multi-agent collaboration into the idea generation process, utilizing LLMs to simulate real-world scientific collaboration scenarios, thereby opening new avenues for generating research ideas. (Li et al., 2024) employed a two-stage approach involving supervised fine-tuning and reinforcement learning to enhance the feasibility, novelty, and effectiveness of research ideas generated by LLMs.(T. Gu et al., 2024) deconstructed paper knowledge into distinct innovative components, leveraging LLMs to combinatorially generate innovative research ideas.

Although existing research has shown that LLMs can produce ideas that are more novel than those written by human experts, it has also highlighted the issue of excessive redundancy in generated ideas (Si et al., 2024) . While Nova (Hu et al., 2024) proposed an iterative planning and search method to reduce the repetition rate of LLM-generated ideas, this study adopts a multi-agent iterative planning and search perspective to further enhance the diversity and novelty of research ideas generated by LLMs.

Prompt Engineering for Logical Reasoning in Large Language Models

Prompt engineering has become an indispensable technique for extending the capabilities of large language models (LLMs) (Sahoo et al., 2024), and the logical reasoning abilities of LLMs are a focal point in the field of artificial intelligence. Consequently, how to leverage prompt engineering to enhance the logical reasoning capabilities of LLMs has become a central focus of scholarly research. Chain-of-Thought (CoT) prompting (Wei et al., 2022) addresses complex problems such as mathematical word problems and commonsense reasoning by presenting reasoning pathways as examples to LLMs, thereby improving their interpretability. Subsequently, (Kojima et al., 2022) proposed Zero-Shot Chain-of-Thought prompting, discovering that simply appending the phrase 'Let's think step by step' to a question enables LLMs to generate a reasoning chain, from which more accurate answers can be extracted. However, creating high-quality Chain-of-Thought examples is time-consuming and labor-intensive. To address this, (Zhang et al., 2022) introduced Auto Chain-of-Thought prompting, which automatically guides LLMs to generate reasoning chains and employs diverse sampling to enhance robustness. (X. Wang et al., 2022) proposed the Self-consistency prompting method, which samples multiple reasoning chains from the LLM's decoder and aggregates them to identify the most consistent answer, significantly improving the performance of Chain-of-Thought methods. Following this, (Zhou et al., 2022)introduced Least-to-Most (LtM)prompting, incorporating planning into prompt engineering by decomposing problems into subproblems and solving them sequentially, thereby enhancing LLMs' ability to tackle complex reasoning tasks. (Yao et al., 2024) proposed the Tree of Thoughts framework, enabling LLMs to explore multiple reasoning paths and self-evaluate before determining the next steps.

While these studies have improved the logical reasoning capabilities of LLMs to some extent, they are limited by the internal knowledge of LLMs and lack interaction with external environments, often leading to hallucinations. To overcome this limitation, (Trivedi et al., 2022) proposed a method combining Chain-of-Thought with external knowledge retrieval, enhancing LLMs' ability to solve knowledge-intensive tasks. Unlike previous approaches that separate reasoning and action in LLMs, ReAct (Yao et al., 2022) allows LLMs to simultaneously generate reasoning trajectories and task-specific actions, fostering synergy between reasoning and action. Specifically, ReAct interacts with external knowledge retrieval tools to address hallucinations and error propagation, thereby improving the factual accuracy of LLM-generated content. In contrast to the linear reasoning chains of LLMs, human thinking is non-linear. To address this, (Besta et al., 2024) introduced Graph of Thoughts prompting, which enables dynamic interaction, backtracking, and evaluation of ideas generated by LLMs, allowing for the aggregation and combination of thoughts from different branches and moving beyond the linear structure of Tree of Thoughts.

Given that a single LLM may be influenced by various biases, leading to inaccuracies in its generated or evaluated outputs (Liusie et al., 2023; P. Wang et al., 2023), many scholars have proposed techniques and architectures for multi-agent LLM systems, such as role-playing (N. Wu et al., 2023), debate (Chan et al., 2023), and voting (Zhu

et al., 2024). Therefore, this study attempts to integrate multi-agent systems with iterative planning to address the complex task of research idea generation.

Combinatorial Innovation Theory and Methodological Approaches

The question of how innovation arises has long been a topic of interest among scholars. Although many theories on innovation are based on human creative activities, they also provide valuable guidance for large language models (LLMs) in engaging in creative endeavors.

Schumpeter proposed that innovation is combinatorial in nature, suggesting that new entities emerge through the recombination of existing elements (Schumpeter & Swedberg, 2021). Boden shares a similar perspective, arguing that novel ideas arise from associating familiar concepts in new ways (Boden, 2004). This mechanism is particularly well-suited for LLMs, which can explore vast knowledge spaces to recombine information and generate novel outputs (T. Gu et al., 2024). In the field of scientometrics, researchers have already begun to explore the application of combinatorial innovation in scientific contexts (Lee et al., 2015; Shi & Evans, 2023; Uzzi et al., 2013) . However, the innovation process is not linear but rather cyclical and iterative, often involving continuous "generation-evaluation" loops (Sharpies, 2013). (Sadler-Smith, 2015) further divides the creative process into four stages—preparation, incubation, insight, and verification—providing a new perspective for understanding creativity.

Combinatorial innovation theory and methodologies offer critical guidance for the design of the approach proposed in this study: A Step-by-Step Research Idea Generation Process: Initial ideas are conceptualized, iteratively refined, and finally summarized to deepen and perfect research directions. Systematic Cross-Domain Knowledge Exploration: A planning-based approach is employed to extensively search knowledge across different domains, using combinatorial knowledge prompts to leverage LLMs' ability to integrate diverse information. Multi-Agent Simulated Brainstorming and Evaluation Mechanism: A multi-agent system simulates the Delphi method to conduct brainstorming sessions, where each agent proposes ideas and an evaluator iteratively assesses them, generating research ideas of greater value.

Data and Methodology

This section provides a detailed exposition of the entire workflow of the multi-agent iterative planning and search strategy. The process comprises four key steps: (1) Dataset Construction; (2) Initial Research Idea Generation;(3) Iterative Refinement of Ideas;(4) Abstract Generation. The framework of the proposed methodology is illustrated in Figure 1.

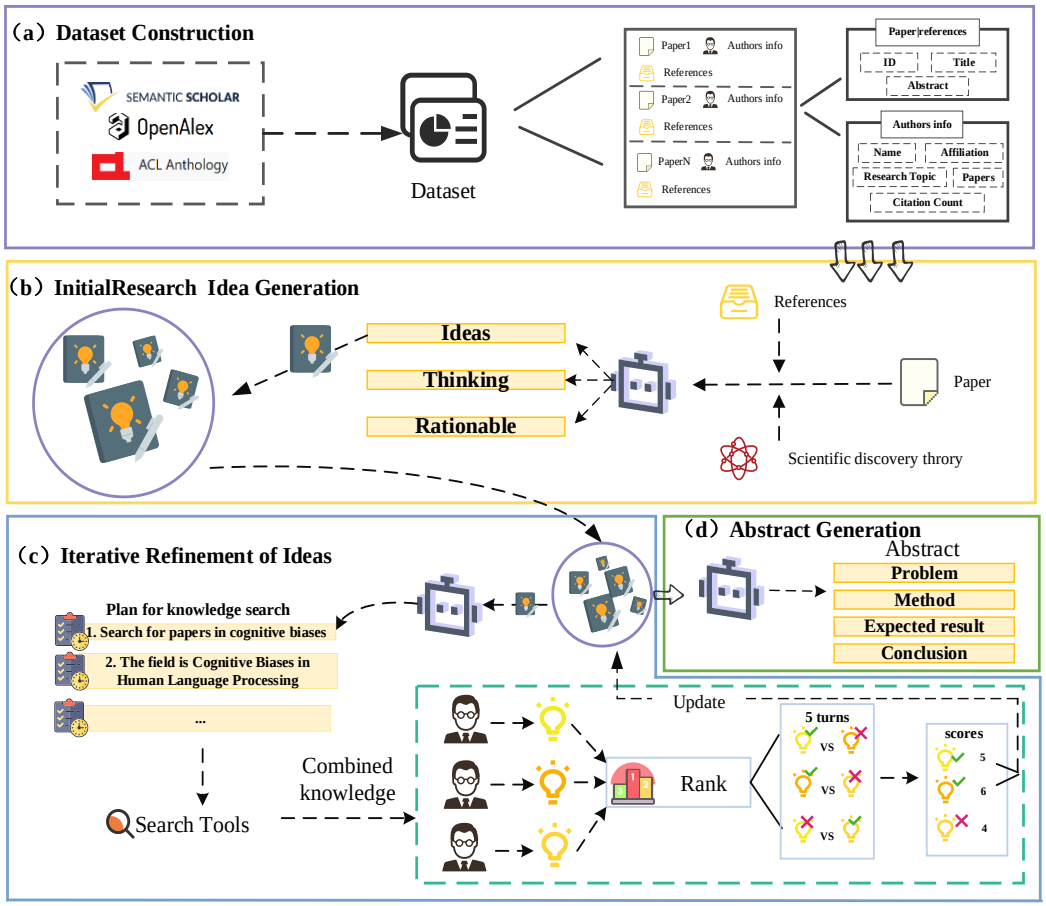


Figure 1. Framework of this study.

Dataset Construction

The data in this study is primarily utilized for two purposes: (1) the generation of initial research ideas and (2) the construction of multi-agent background information. To achieve this, the study requires access to target papers, their references, and information about the authors of the target papers. Research by (Guo et al., 2024) has demonstrated that high-quality papers significantly enhance the quality of research ideas generated by large language models (LLMs). Therefore, this study selects long papers from the 2024 Annual Meeting of the Association for Computational Linguistics (ACL)[1] as the initial corpus. However, a single paper database is insufficient to meet the data requirements of this study. Consequently, during the data collection process, this study leverages multiple data sources, including the ACL Anthology Corpus[1], OpenAlex[2](Priem et al., 2022) and Semantic Scholar[3](Kinney et al., 2023), to gather the necessary data.

Ultimately, we successfully collected 675 target papers along with their corresponding 22,647 references. To ensure data quality, further filtering was

[1] https://aclanthology.org/
[2] https://openalex.org/
[3] https://www.semanticscholar.org/

applied to exclude target papers with fewer than 10 citations, fewer than 20 references, or missing author information. After rigorous screening, the final dataset consists of 144 target papers, 6,153 references, 953 author profiles, and 25,906 papers published by the corresponding authors.

In this dataset, the data fields for the target papers and their references include the paper titles and abstracts. The author information fields encompass research interests, affiliated institutions, publication counts, citation counts, and the papers they have published. Additionally, to protect privacy, sensitive information such as names in the dataset has been appropriately anonymized.

Initial Research Idea Generation

This study begins by randomly selecting a target paper to define the direction for research idea generation and to determine the scale of the multi-agent team. Drawing on(Sadler-Smith, 2015) framework, which divides the creative process into four stages—preparation, incubation, insight, and verification—the initial research idea generation phase aims to prepare and incubate ideas, laying the groundwork for subsequent iterations by the agents to produce truly novel ideas. To this end, an initial idea generation module is designed, emphasizing diversity and novelty as foundational principles. Upon receiving the input paper, the large language model (LLM) utilizes its references and scientific discovery theories to generate ideas.

To enhance the scientific rigor and diversity of the initial research ideas, this study adopts an approach inspired by Nova(Hu et al., 2024) , employing ten scientific discovery methods to constrain and stimulate the LLM. These methods guide the LLM to generate innovative ideas based on the input paper and its references. For example, leveraging Pierce's hypothetico-deductive method, the model starts with facts and propositions, formulates a hypothesis or premise, and then conducts logical reasoning to derive conclusions. By analyzing the relationships between premises, the validity and truth value of the conclusions can be assessed.

In alignment with the creative process, the study utilizes the internal knowledge of the LLM to stimulate idea generation, ensuring that the model comprehends the input paper and its references, evaluates them, and provides reasoning and thought processes to maintain interpretability(Wei et al., 2022). Finally, 15 initial research ideas are generated, forming an idea pool to facilitate subsequent iterations.

To formalize the prompting process, this study defines P as the target paper, L as its references, T as the scientific method theory, and R as the generated research idea. Thus, the initial research idea generation can be expressed as:

$$R = f(P,L,T) \tag{1}$$

Where f represents the large language model, leveraging its language comprehension capabilities to generate research ideas. The prompt templates and examples for initial research idea generation are provided in Appendix Tables 1 and 2.

Iterative Refinement of Ideas

Previous methods have predominantly relied on keyword-based searches or co-occurrence of entity concepts to incorporate external knowledge. However, these approaches exhibit significant limitations, such as inaccurate or overly broad search results, which hinder the ability of large language models (LLMs) to engage in deep reasoning (Hu et al., 2024).

To effectively address these shortcomings, this study integrates planning principles into the knowledge search phase of research idea generation. Specifically, the LLM is utilized to meticulously plan and design knowledge search tasks, which are then executed sequentially using external academic search APIs. Ultimately, knowledge from diverse domains that is closely related to the research idea is combinatorially integrated into the LLM's prompts, providing more targeted and novel composite knowledge for idea generation. The prompt templates and examples for knowledge planning and search are provided in Appendix Tables 3 and 4.4.

The multi-agent system constructed in this study comprises multiple agents, each endowed with background knowledge of real-world scientists, denoted as $S = [s_1, s_2, ..., s_n]$, where represents the entire scientific agent team and n 为 denotes the team size. The background knowledge of these scientific agents is derived from the author team information of the target papers. In the iterative process of research idea generation, these agents simulate the Delphi method, a widely recognized practice in innovation. Specifically, upon acquiring new knowledge, each agent proposes its own research ideas and conducts self-evaluation and scoring based on best practices from AI conference reviews (e.g., ICLR and ACL) (Si et al., 2024). The scoring criteria are provided to each agent as contextual prompts. Detailed scoring guidelines can be found in Appendix Table 5.

The research ideas generated by each scientific agent are evaluated for their creative quality using a Swiss System Tournament and a zero-shot large language model (LLM) ranker. The ranker employs a pairwise comparison approach to determine which idea is superior. Each idea undergoes five rounds of comparison, with a score of 1 point awarded for each win. Empirical evidence suggests that this quality assessment method outperforms direct comparison approaches(Lu et al., 2024) . Ultimately, ideas scoring 5 points or higher are selected as the final output of the current iteration. Additionally, the negative feedback recorded during the comparison process is carried forward, along with the selected final ideas, into the next iteration. The prompting process for each scientific agent can be expressed as:

$$\mathrm{R}_i = f(R_t, K, B) \tag{2}$$

Where, R_i represents the research idea generated by the i-th scientific agent, R_t denotes the research idea generated in the t-th iteration, K signifies the new knowledge acquired through planning and search, and B represents the feedback from the research ideas generated in the t-th iteration. The prompt templates and examples for research idea generation can be found in Appendix Tables 6 and 7, while the prompts for research idea comparison are provided in Appendix Table 8.

In each iteration, newly generated research ideas replace the older ones. Through this mechanism, the agents in this study are able to conduct more in-depth research exploration, significantly expanding the boundaries of the search space.

Abstract Generation

After T iterations, the final research ideas are established. In this process, the study draws on the summary generation method proposed by VIRSCI (Su et al., 2024). Specifically, the finalized research ideas are input into the large language model (LLM) with a rigorously defined summary format (including aspects such as objectives and problems, methods, expected results, and conclusions), ensuring that the research ideas are presented in a detailed and structured manner. Additionally, since the summaries will subsequently be compared with reference paper abstracts for evaluation, outputting the research ideas in summary form is both practical and aligned with the assessment requirements. The prompt templates and examples for research idea summary generation can be found in Appendix Tables 9 and 10.

Experiments and Results Analysis

This section conducts comprehensive experiments to evaluate the effectiveness of the multi-agent iterative planning and search strategy, followed by an in-depth analysis and interpretation of the results.

Experimental Setup

Large Language Model Configuration

We implements the proposed method within the multi-agent application framework Agentscope[4](Gao et al., 2024). The large language model (LLM) employed in this study is DeepSeek-V3[5] , which has demonstrated superior performance across multiple benchmarks compared to other open-source models such as Qwen2.5-72B[6] and Llama-3.1-405B[7]. Additionally, its performance is on par with world-leading proprietary models, including GPT-4o [8] and Claude-3.5-Sonnet[9] (Liu et al., 2024).

Baselines

To demonstrate the effectiveness of the proposed method, this study selects state-of-the-art approaches as baselines, including AI-Researcher[10](Si et al., 2024). This method introduces an end-to-end framework for generating research ideas using large language models (LLMs) and demonstrates that LLM-generated ideas are more novel than those produced by human experts.

[4] https://github.com/modelscope/agentscope
[5] https://platform.deepseek.com/
[6] https://huggingface.co/Qwen/Qwen2.5-72B
[7] https://huggingface.co/meta-llama/Llama-3.1-405B
[8] https://openai.com/index/hello-gpt-4o/
[9] https://www.anthropic.com/claude/sonnet
[10] https://github.com/NoviScl/AI-Researcher

Evaluation Metrics

Drawing on the evaluation methodologies of AI-Researcher(Si et al., 2024) and Nova(Hu et al., 2024) , We assesses the research ideas generated by LLMs from three perspectives: quality score, diversity, and novelty.

(1) Quality Score**:** The quality of research ideas is evaluated using a Swiss System Tournament and a zero-shot LLM ranker. Specifically, the ranker employs a pairwise comparison approach to determine which idea is superior. Each idea undergoes five rounds of comparison, with 1 point awarded for each win. This quality assessment method has been empirically shown to outperform direct comparison or scoring approaches (Lu et al., 2024) . Ideas scoring above 5 points are considered high-quality. The quality score is ultimately measured by the proportion of high-quality ideas, calculated as follows:

$$HightScoreRatio = \frac{\sum_{i}^{n} I(s_i \geq 5)}{n} \tag{3}$$

Where, n represents the total number of generated research ideas, s_i denotes the score of the i-th idea, and $I(s_i \geq 5)$ is an indicator function that equals 1 when $s_i \geq 5$ and 0 otherwise.

(2) Novelty: We employ semantic similarity to assess the novelty of research ideas generated by large language models. Specifically, we first use a text embedding model to convert the generated research ideas and relevant literature into vector representations, then calculate the similarity between them. If the similarity falls below a predefined threshold, the idea is considered novel. This approach has been widely adopted in the evaluation of research idea generation(Hu et al., 2024; Kumar et al., 2024; Si et al., 2024). Additionally, the all-MiniLM-L6-v2[11] model is used for embedding, with a cosine similarity threshold of 0.5 to determine similarity. The novelty score is calculated as follows:

$$Novelty = \frac{\sum_{i=1}^{n} I(max_j sim(a_i, r_{ij}) < \theta)}{n} \tag{4}$$

Where, n represents the total number of generated research ideas, $sim(a_i, r_{ij})$ denotes the cosine similarity between the i-th idea a_i and its related literature r_{ij}, $I(\cdot)$ is an indicator function that returns 1 if the condition is true and 0 otherwise.

(3) Diversity: Similar to(Hu et al., 2024; Si et al., 2024) , the diversity of generated research ideas is measured by the proportion of unique ideas. Specifically, the same similarity metric used for novelty assessment is applied, with a duplication threshold set at 0.8. The diversity score is calculated as follows:

[11] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

$$Diversity = \frac{1}{n}\sum_{i=1}^{n} I(max_{j\neq i} sim(i,j) < threshold) \quad (5)$$

Where, n represents the total number of generated research ideas, $sim(i,j)$ 是 denotes the cosine similarity between the i-th idea and j-th idea, $threshold$ represents the similarity threshold, and $I(\square)$ is an indicator function that returns 1 if the condition is true and 0 otherwise.

During the evaluation process, this method randomly selects 5 papers for each team size ranging from 2 to 8 members, totaling 35 papers, and generates 525 initial research ideas. Each baseline method also produces 5 sets of data, resulting in 75 research ideas for evaluation. The final assessment is based on the average proportion of high-quality ideas, average novelty, and average diversity scores.

Throughout the experiments, the multi-agent iterative planning and search strategy, after three iterations, cumulatively generated 2,027 research ideas. Specifically, the first iteration produced 568 ideas, the second iteration generated 656 ideas, and the third iteration yielded 803 ideas. In detail, teams of 8 members contributed 126 ideas, 7-member teams generated 113 ideas, 6-member teams produced 107 ideas, 5-member teams contributed 97 ideas, 4-member teams formed 79 ideas, 3-member teams created 77 ideas, and 2-member teams generated 75 ideas. The trend in the average number of ideas per team size is illustrated in Figure 2. Clearly, larger team sizes result in a greater number of research ideas after filtering through the LLM's self-evaluation process.

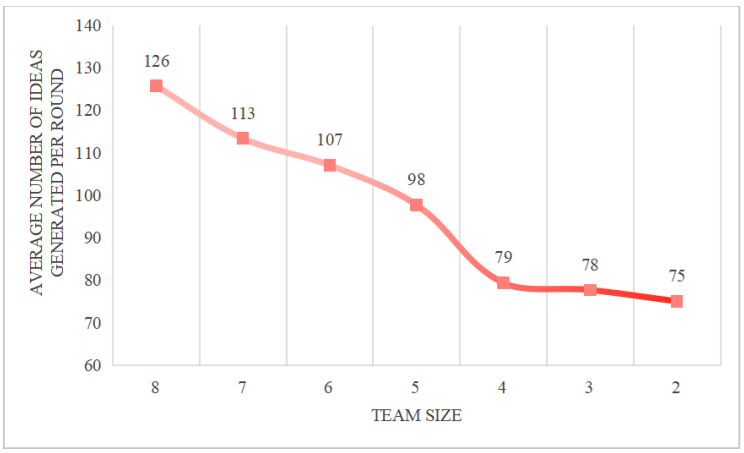


Figure 2. Average Number of Ideas Generated per Team per Iteration.

Comparison with Baseline Methods

We answer RQ1 in this section. Following the methodology of AI-Researcher (Si et al., 2024) , we fully replicate their approach and, to ensure consistency with our method, generate 5 sets of data to obtain 75 research ideas for analysis. For our proposed method, we use the average performance metrics across iterations to ensure

a fair comparison. The results are presented in Figure 3. Our method outperforms AI-Researcher in both the average diversity ratio and the proportion of high-quality ideas, while also demonstrating a slight advantage in the average novelty ratio. These findings indicate that the multi-agent iterative planning and search strategy can effectively enhance the diversity and novelty of research ideas generated by large language models.

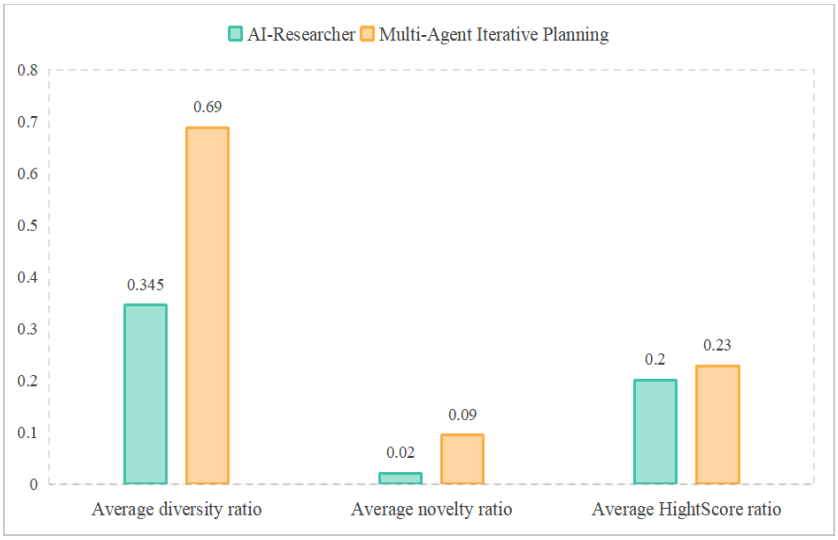


Figure 3. Comparison with Baseline Methods.

Impact of Agent Team Size on Performance Metrics

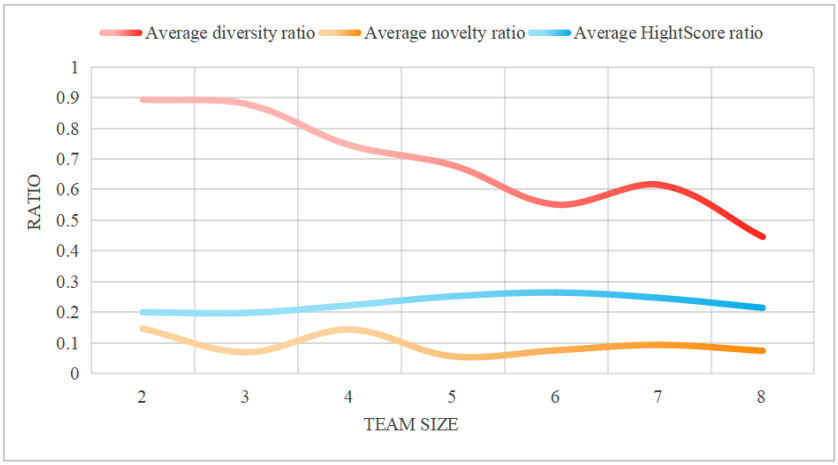


Figure 4. Trend of Metrics Across Different Team Sizes.

We examine the impact of varying agent team sizes on performance metrics by analyzing the best-performing third iteration results. As shown in Figure 4, for diversity, as the team size increases from 2 to 8, the uniqueness ratio exhibits an overall declining trend, starting from a relatively high level and gradually decreasing.

This suggests that larger team sizes may lead to a reduction in uniqueness, which is likely related to the inherent knowledge limitations of large language models (LLMs). Generating more content increases the likelihood of similarity, indicating that expanding the scale of multi-agent systems does not necessarily enhance the uniqueness of LLM-generated content. This reflects a trade-off between quality and uniqueness.

For novelty, no clear trend is observed in relation to team size. However, the overall values remain relatively low and stable, indicating that team size has an insignificant impact on novelty. This further suggests that the proposed method cannot improve novelty by scaling up the number of agents.

The proportion of high-quality ideas fluctuates between 0.2 and 0.3 as the team size varies from 2 to 8, without showing a clear linear increase or decrease. However, in local variations:

Small teams (team size of 2-3): The proportion of high-quality ideas is relatively low, around 0.2. This may be due to limited resources and manpower in smaller teams, making it difficult to achieve high performance across all aspects.

Medium-sized teams (team size of 4-7): The proportion of high-quality ideas increases and stabilizes around 0.25. At this scale, teams may achieve a better balance in personnel allocation and collaboration, leading to improved overall performance.

Large teams (team size of 8): The proportion of high-quality ideas drops back to around 0.2. This may be attributed to increased management complexity and communication costs in larger teams, which can negatively impact overall efficiency and quality.

These findings align with the conclusion that an optimal team size can facilitate the generation of impactful research (L. Wu et al., 2019).

Impact of Iteration Count on Performance Metrics

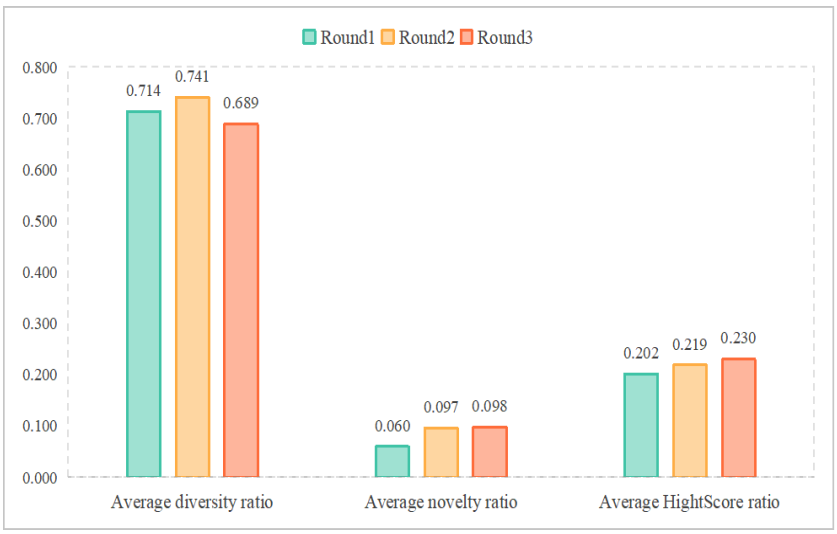


Figure 5. Impact of Iteration Count on Average Metrics.

As shown in Figure 5, the number of iterations has a significant impact on all metrics. The average diversity ratio peaks during the second iteration and then slightly declines. The average novelty ratio shows a notable improvement in the second iteration, with a marginal increase in the third iteration. Meanwhile, the proportion of high-quality ideas gradually rises with each iteration. These results suggest that the proposed method retains potential for generating high-quality ideas, though it exhibits some limitations in terms of novelty and diversity.

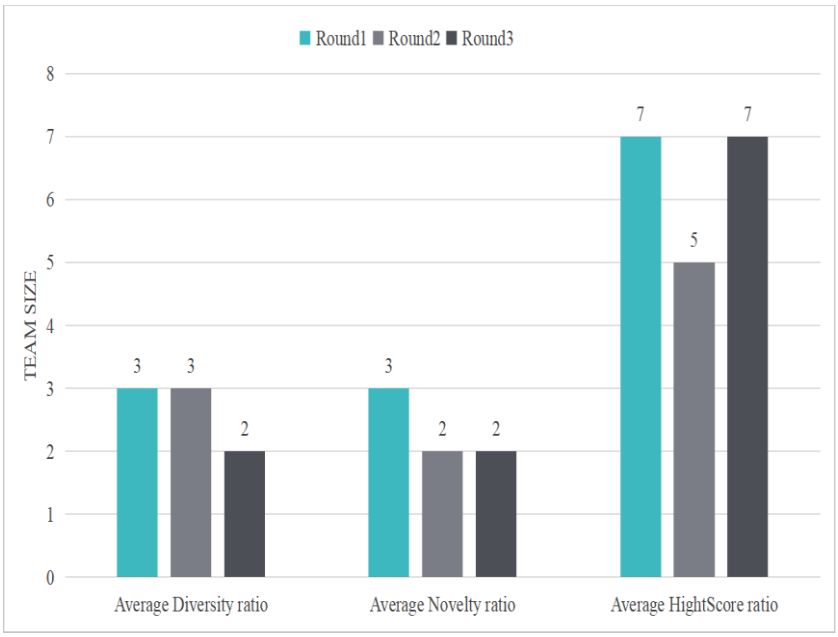


Figure 6. Variation in Team Size Corresponding to the Best Metrics per Iteration.

Across different iteration counts, as illustrated in Figure 6, the best performance in diversity and novelty metrics consistently occurs in smaller teams, while the highest proportion of high-quality ideas is consistently achieved by teams of 5-7 members. This indicates that the multi-agent strategy holds promise for enhancing the quality of research ideas generated by large language models.

Ablation Study

Figure 7. Comparative Performance of Single-Agent vs. Multi-Agent Systems Across Iterations.

We answer RQ2 in this section. The multi-agent iterative planning and search strategy integrates two core modules: knowledge planning and search, and multi-agent generation. A key objective of this study is to determine which module plays

a decisive role in influencing critical metrics. To this end, we set the number of agents to 1, focusing on the impact of a single agent combined with knowledge planning and search on research idea generation. The best performance of the multi-agent iterative planning and search strategy is used as a benchmark for comparison.
As shown in Figures 7(a), 7(b), and 7(c), the single-agent approach outperforms in terms of diversity and novelty metrics. This suggests that the knowledge planning and search module positively enhances the generative capabilities of large language models (LLMs), indicating that combinatorial knowledge effectively guides LLMs. However, we also observe a declining trend in the performance of the single-agent system as the experiment progresses, suggesting that it may encounter bottlenecks in the research idea generation process. This finding aligns with the conclusions of Hu et al. (2024) in their study on Nova, where performance similarly plateaued after a certain number of iterations.
In contrast, while the multi-agent system slightly underperforms in diversity and novelty compared to the single-agent approach, it demonstrates significant advantages in the quality of generated ideas. Notably, the multi-agent system exhibits a consistent upward trend across all metrics. This indicates the potential of multi-agent systems and highlights the feasibility of incorporating innovative methodologies. In other words, combinatorial innovation theory and methodological approaches can effectively guide LLMs in the task of generating research ideas.

Discussion

Research Implications

Theoretical Implications

This study applies combinatorial innovation theory to the task of research idea generation, proposing a novel methodological framework aimed at enhancing the diversity and novelty of research ideas generated by large language models (LLMs). Experimental results demonstrate that the proposed method consistently outperforms baseline approaches across key evaluation metrics, including diversity, novelty, and quality scores. This underscores the effectiveness of systematically combining knowledge from diverse domains and employing multi-agent systems to conduct 'brainstorming' sessions. Furthermore, it validates the feasibility of applying combinatorial innovation theory and practical innovation methodologies to the task of research idea generation.
In the ablation study, we compared the individual contributions of the knowledge planning module and the multi-agent system module. The findings reveal that the knowledge planning and search module positively influences the generative capabilities of LLMs, confirming that combinatorial knowledge effectively guides LLMs. Additionally, the multi-agent system, unlike the single-agent approach, maintains an upward performance trend over more iterations, suggesting that practical innovation methodologies can effectively guide LLMs in performing complex reasoning tasks. In summary, combinatorial innovation theory and methodologies are well-suited for the task of research idea generation.

Practical Implications

The multi-agent iterative planning and search strategy significantly enhances the performance of research idea generation tasks in terms of diversity, novelty, and quality. This suggests that adopting a collaborative approach involving multiple modules or agents often yields superior outcomes when conducting research idea generation tasks. Furthermore, leveraging theoretical frameworks to guide the design of each module is crucial. Such an approach not only ensures a more scientific practice but also enhances the interpretability of the results, facilitating a deeper understanding and application of the research findings.

Limitations

In this study, we employ knowledge planning and search alongside a multi-agent system to simulate human innovation processes, aiming to enhance the innovative capabilities of large language models (LLMs) in generating research ideas. Despite promising results, this work has several limitations.

Incomplete Evaluation Metrics: Although the evaluation metrics used in this study encompass novelty, diversity, and quality comparison, they do not account for the value, feasibility, or historical impact of the generated research ideas. This may limit the applicability of the findings.

Lack of Human Expert Evaluation: Although the automated evaluation results indicate that the proposed method in this paper excels in enhancing the novelty and diversity of research ideas generated by large language models, relying solely on automated metrics makes it difficult to comprehensively validate the method's reliability. In subsequent research, we will incorporate human expert evaluation to further verify the practical effectiveness of the method.

Cross-disciplinary applicability remains to be verified: Although this study has demonstrated the effectiveness of multi-agent iterative search strategies for generating research ideas with large language models in the field of natural language processing (NLP), the success in a single discipline is insufficient to prove the universality of this approach. This limitation may hinder its broader application in other subfields of computer science or even more extensive disciplines (such as life sciences, physics, etc.).

Absence of Reward Functions: While the process of generating research ideas with LLMs incorporates combinatorial knowledge from multiple domains, it relies solely on the inherent capabilities of the model without introducing reward mechanisms to guide the generation process. This could potentially impact the quality of the generated ideas.

Conclusion and Future Research Directions

This study introduces combinatorial innovation theory into the task of research idea generation and proposes a multi-agent iterative planning and search strategy that integrates multi-domain knowledge planning and search with a multi-agent system. Experimental results demonstrate that this method outperforms baseline approaches, enhancing the diversity and novelty of generated research ideas. Furthermore, it provides a theoretical explanation, grounded in combinatorial innovation theory and

methodologies, for why the proposed method improves the diversity and novelty of ideas generated by language models, offering new perspectives for future research on idea generation tasks.

Future research efforts will focus on three key directions: (1) Enhancing the research idea generation capability of large language models (LLMs) through fine-grained knowledge entity recombination techniques;(2) Establishing a multi-dimensional evaluation framework to systematically validate the academic value and practical effectiveness of generated content;(3) Constructing domain-specific knowledge graphs to constrain and guide LLM generation processes, thereby effectively mitigating hallucination phenomena.

Ethical Statement

All literature data and author background information used in this study were sourced from publicly available academic databases, and none of the content involves personal privacy or sensitive information. During data processing, we strictly adhered to the terms of use and academic ethics guidelines of each database to ensure no risk of privacy breaches. To protect scholars' personal information security, all author names were anonymized during the analysis.

It is important to emphasize that the system developed in this study is solely intended to assist scientific research. Its design purpose is to provide research support for scholars, not to replace human researchers. Throughout its operation, the system emphasizes the importance of human oversight mechanisms, ensuring the quality of research results through human-machine collaboration.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Grant No.72074113).

References

Ajith, A., Xia, M., Chevalier, A., Goyal, T., Chen, D., & Gao, T. (2024). Litsearch: A retrieval benchmark for scientific literature search. *arXiv preprint arXiv:2407.18940*.

Baek, J., Jauhar, S. K., Cucerzan, S., & Hwang, S. J. (2024). Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Podstawski, M., Gianinazzi, L., Gajda, J., Lehmann, T., Niewiadomski, H., & Nyczyk, P. (2024). *Graph of thoughts: Solving elaborate problems with large language models*. Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence.

Boden, M. A. (2004). *The creative mind: Myths and mechanisms*: Routledge.

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2023). Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.

Gao, D., Li, Z., Pan, X., Kuang, W., Ma, Z., Qian, B., Wei, F., Zhang, W., Xie, Y., & Chen, D. (2024). Agentscope: A flexible yet robust multi-agent platform. *arXiv preprint arXiv:2402.14034*.

Gu, T., Wang, J., Zhang, Z., & Li, H. (2024). LLMs can realize combinatorial creativity: generating creative ideas via LLMs for scientific research. *arXiv preprint arXiv:2412.14141*.

Gu, X., & Krenn, M. (2024). Generation and human-expert evaluation of interesting research ideas using knowledge graphs and large language models. *arXiv preprint arXiv:2405.17044*.

Guo, S., Shariatmadari, A. H., Xiong, G., Huang, A., Xie, E., Bekiranov, S., & Zhang, A. (2024). IdeaBench: Benchmarking Large Language Models for Research Idea Generation. *arXiv preprint arXiv:2411.02429*.

Hu, X., Fu, H., Wang, J., Wang, Y., Li, Z., Xu, R., Lu, Y., Jin, Y., Pan, L., & Lan, Z. J. a. p. a. (2024). Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., & Cohan, A. (2023). The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems, 35*, 22199-22213.

Kumar, S., Ghosal, T., Goyal, V., & Ekbal, A. (2024). Can Large Language Models Unlock Novel Scientific Research Ideas? *arXiv preprint arXiv:2409.06185*.

Larivière, V., Archambault, É., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and technology, 59*(2), 288-296.

Lee, Y.-N., Walsh, J. P., & Wang, J. J. R. p. (2015). Creativity in scientific teams: Unpacking novelty and impact. *44*(3), 684-697.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems, 33*, 9459-9474.

Li, R., Jing, L., Han, C., Zhou, J., & Du, X. (2024). Learning to Generate Research Idea with Dynamic Control. *arXiv preprint arXiv:2412.14626*.

Linstone, H. A., & Turoff, M. (1975). *The delphi method*: Addison-Wesley Reading, MA.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., & Ruan, C. (2024). Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liusie, A., Manakul, P., & Gales, M. J. (2023). Zero-shot nlg evaluation through pairware comparisons with llms. *arXiv preprint arXiv:2307.07889*.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Pu, K., Feng, K., Grossman, T., Hope, T., Mishra, B. D., Latzke, M., Bragg, J., Chang, J. C., & Siangliulue, P. (2024). IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. *arXiv preprint arXiv:2410.04025*.

Sadler-Smith, E. J. C. r. j. (2015). Wallas' four-stage model of the creative process: More than meets the eye? , *27*(4), 342-352.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Liu, Z., & Barsoum, E. (2025). Agent Laboratory: Using LLM Agents as Research Assistants. *arXiv preprint arXiv:2501.04227*.

Schumpeter, J. A., & Swedberg, R. (2021). *The theory of economic development*: Routledge.

Sharpies, M. (2013). An account of writing as creative design *The science of writing* (pp. 127-148): Routledge.

Shi, F., & Evans, J. J. N. C. (2023). Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *14*(1), 1641.

Si, C., Yang, D., & Hashimoto, T. (2024). Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *rXiv preprint arXiv:.2409.04109*.

Su, H., Chen, R., Tang, S., Zheng, X., Li, J., Yin, Z., Ouyang, W., & Dong, N. (2024). Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *arXiv preprint arXiv:2410.09403*.

Trivedi, H., Balasubramanian, N., Khot, T., & Sabharwal, A. (2022). Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. J. S. (2013). Atypical combinations and scientific impact. *342*(6157), 468-472.

Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems, 35*, 24824-24837.

Wu, L., Wang, D., & Evans, J. A. J. N. (2019). Large teams develop and small teams disrupt science and technology. *566*(7744), 378-382.

Wu, N., Gong, M., Shou, L., Liang, S., & Jiang, D. (2023). *Large language models are diverse role-players for summarization evaluation*. Paper presented at the CCF International Conference on Natural Language Processing and Chinese Computing.

Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R. J., & Anandkumar, A. (2024). Leandojo: Theorem proving with retrieval-augmented language models. *Advances in Neural Information Processing Systems, 36*.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems, 36*.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Yuan, J., Yan, X., Shi, B., Chen, T., Ouyang, W., Zhang, B., Bai, L., Qiao, Y., & Zhou, B. (2025). Dolphin: Closed-loop Open-ended Auto-research through Thinking, Practice, and Feedback. *arXiv preprint arXiv:2501.03916*.

Zhang, Z., Zhang, A., Li, M., & Smola, A. J. a. p. a. (2022). Automatic chain of thought prompting in large language models.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., & Le, Q. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Zhu, K., Wang, J., Zhao, Q., Xu, R., & Xie, X. (2024). *Dynamic Evaluation of Large Language Models by Meta Probing Agents*. Paper presented at the Forty-first International Conference on Machine Learning.

Table 1. Prompt for initial Research Idea Generation.

System prompt: You are an expert researcher in AI. Your goal is to propose some innovative and valuable research ideas based on the target paper.
Follow these steps to generate innovative research ideas for exploration: Understand the Target Paper and Related Works: Target Paper: This is the core research study you aim to enhance or build upon. It serves as the foundation for identifying and developing new research ideas. Referenced Papers: These are studies cited by the target paper, providing additional context and insights directly relevant to the primary research topic. They are crucial for understanding and expanding upon the target paper. Leverage Scientific Discovery Theories: Select appropriate scientific discovery theories and combine them with insights from the target paper to generate creative, impactful, and feasible research ideas. Explore Scientific Discovery Methodologies: Below are 10 general laws and methodologies of scientific discovery from the philosophy of science. Choose one or more of these methodologies to propose new research ideas for the target paper: {scientific_discovery_theory} Select and Propose New Ideas: Identify the 5 most suitable theories or methods for the target paper and propose 5 new research ideas based on them. Requirements: Output: Provide 5 new research ideas worth exploring. Quality: Aim for ideas with the potential to win best paper awards at top conferences like ACL, NeurIPS, ICLR, and CVPR. Relevance: Skip theories that do not align well with the target paper. Ensure the chosen theories and methods are logical and appropriate for the research context. Input: Target Paper: {target_paper} References: {references} Response Format: Thought: <THOUGHT> IDEA: ```json<JSON> ``` In <THOUGHT>, explain your thought process, including why you chose specific theories and why the proposed ideas have the potential to win best paper awards at top conferences. In <JSON>, provide the new idea with the following fields: Title: A concise title for the idea, suitable for a research paper. Idea: A detailed description of the idea, including its significance and potential impact.

Thinking: A detailed explanation of the thought process, including the scientific discovery theory used. Rationale: A detailed justification for why this idea is innovative and feasible. Ensure the JSON format is precise, as it will be automatically parsed.

Table 2. An example for initial research idea generation.

Title: Beyond Linguistic Features: Leveraging Cognitive Biases for Machine-Generated Text Detection
Idea: This idea proposes a novel approach to detecting machine-generated text by leveraging cognitive biases inherent in human language processing. While current methods focus on linguistic features, this approach hypothesizes that machine-generated text lacks subtle cognitive biases that humans exhibit, such as confirmation bias, anchoring, and availability heuristic. By training models to detect these biases, we can create a more robust detector that performs well even when linguistic distinctions are minimal. The idea involves collecting a dataset of human and machine-generated texts annotated for cognitive biases, developing a model to detect these biases, and evaluating its performance in diverse scenarios.
Thinking: This idea is inspired by Pierce\u2019s hypothetical deduction method and Kuhn\u2019s paradigm theory. The hypothesis is that machine-generated text lacks cognitive biases, which are subtle but consistent features of human language. By shifting the paradigm from linguistic features to cognitive biases, we can address the anomaly of decreasing linguistic distinctions. The rationale is that while machines can mimic linguistic patterns, they cannot replicate the nuanced cognitive biases that humans naturally exhibit.
Rationale: Current detection methods struggle with out-of-distribution texts and are limited by the decreasing linguistic distinctions between human and machine-generated text. By focusing on cognitive biases, which are deeply ingrained in human language, we can create a more robust detection method that is less reliant on surface-level linguistic features. This approach has the potential to significantly improve detection accuracy in real-world scenarios, making it a strong candidate for a best paper award.

Table 3. Prompt for knowledge planning and search.

System prompt: You are now a senior scholar in the field of artificial intelligence, and you need to develop a search strategy based on a given research idea
Objective: To analyze a given research idea and determine which fields of papers should be retrieved to collect comprehensive information and new knowledge. This will provide a basis for further research and the discovery of new ideas. Input: Research Idea: {idea} Response Format: Thought: <THOUGHT> Search Plans: ```json<JSON> ```

In <THOUGHT>, explain the reasoning behind the search strategy, including why specific fields and keywords were chosen.

In <JSON>, provide a detailed search plan with the following fields:

Search Plan: A brief description of the overall search strategy.

Search Fields: A list of fields relevant to the research idea.

Search Keywords: A structured list of keywords for each field, formatted as:

"Search_Keywords": [

{"Field": "field_name", "Keywords": ["keyword1", "keyword2", "keyword3"]}

]

Ensure the JSON format is precise, as it will be automatically parsed.

Table 4. An example for knowledge planning and search.

Search Plan: "To gather comprehensive information, we will search for papers in cognitive biases, machine-generated text detection, NLP, dataset creation, and model evaluation. Each field will be explored using specific keywords to ensure relevance and depth."
Search Keyword": [{"Field": "Cognitive Biases in Human Language Processing", "Keywords": ["confirmation bias", "anchoring bias", "availability heuristic", "cognitive biases in language", "human language processing"]},{"Field": "Machine-Generated Text Detection", "Keywords": ["machine-generated text", "AI-generated text", "text detection", "linguistic features", "deepfake text"]},{"Field": "Natural Language Processing (NLP)","Keywords": ["text analysis", "language models", "NLP techniques", "text classification", "semantic analysis"]}.
Thought: The research idea revolves around detecting machine-generated text by leveraging cognitive biases in human language processing. To develop a comprehensive search strategy, we need to explore several interconnected fields: (1) **Cognitive Biases in Human Language Processing** to understand the theoretical foundation of the biases, (2) **Machine-Generated Text Detection** to review existing methods and their limitations, (3) **Natural Language Processing (NLP)** for techniques to analyze and model text, (4) **Dataset Creation and Annotation** for methodologies to collect and label data, and (5) **Model Evaluation** for metrics and frameworks to assess detector performance. This strategy ensures a holistic understanding of the problem and provides a foundation for developing a novel approach.

Table 5. Scoring rubric.

Novelty：Whether the idea is creative and different from existing works on the topic, and brings fresh insights. You are encouraged to search for related works online. You should consider all papers that appeared online prior to 2024 as existing work when judging the novelty. A rating from 1 to 10.Here are the grading rules: 1. Not novel at all - there are many existing ideas that are the same

2.
3. Mostly not novel - you can find very similar ideas
4.
5. Somewhat novel - there are differences from existing ideas but not enough to turn into a new paper
6. Reasonably novel - there are some notable differences from existing ideas and probably enough to turn into a new paper
7.
8. Clearly novel - major differences from all existing ideas
9.
10. Very novel - very different from all existing ideas in a very interesting and clever way

Feasible: How feasible it is to implement and execute this idea as a research project? Specifically, how feasible the idea is for a typical CS PhD student to execute within 1-2 months of time. You can assume that we have rich API resources, but only limited hardware resources. A rating from 1 to 10.Here are the grading rules:
1. Impossible: the idea doesn't make sense or the proposed experiments are flawed and cannot be implemented
2.
3. Very challenging: there are flaws in the proposed method or experiments, or the experiments require compute/human resources beyond any academic lab
4.
5. Moderately feasible: It can probably be executed within the given time frame but would require careful planning, efficient use of APIs or some advanced computational strategies to overcome the limited GPU resources, and would require some modifications to the original proposal to make it work
6. Feasible: Can be executed within the given constraints with some reasonable planning
7.
8. Highly Feasible: Straightforward to implement the idea and run all the experiments
9.
10. Easy: The whole proposed project can be quickly executed within a few days without requiring advanced technical skills

Excitement: How exciting and impactful this idea would be if executed as a full project. Would the idea change the field and be very influential. A rating from 1 to 10.Here are the grading rules:
1. Poor: You cannot identify the contributions of this idea, or it's not interesting at all and you would fight to have it rejected at any major AI conference
2.
3. Mediocre: this idea makes marginal contributions and is very incremental
4.
5. Leaning negative: it has interesting bits but overall not exciting enough

6. Learning positive: exciting enough to be accepted at a major AI conference, but still has some weaknesses or somewhat incremental 7. 8. Exciting: would deepen the community's understanding or make major progress in this research direction 9. 10. Transformative: would change the research field profoundly and worth a best paper award at major AI conferences

Note: Some score values in the scoring rubric lack descriptions. This is because the granularity of the score levels is challenging to articulate in English. For specific details, please refer to the approach used in AI-Researcher[12].

Table 6. Prompt for research idea generation.

System prompt: Your name is Scientist0, you belong to following affiliations ['Westlake University'], you have researched on following topics ['Natural Language Processing Techniques', 'Topic Modeling', 'Multimodal Machine Learning Applications', 'Text Readability and Simplification', 'Speech and dialogue systems'], you have published 24 papers, you have 844 citations.
Task: You are an ambitious scientist who is looking to propose a new idea that will contribute significantly to the field. Improve the seed idea or come up with the next impactful and creative idea for publishing a paper that will contribute significantly to the field by integrating your own knowledge and insights with the information provided.
Here is the seed idea : '"{seed_idea}"'
Here are the bad reviews of the seed idea, please improve according to the reviews: '"{bad reviews}"'
You may refer to the following listed references to design a new idea or concept. These references can serve as inspiration, but you are not allowed to directly copy or replicate their content. Ensure that your design is original and addresses a specific problem or meets a unique need. References: {references} "Please respond in the following format: Thought: <THOUGHT> New Idea: ```json<JSON>``` In <THOUGHT>, briefly discuss your intuitions and motivations for the idea. Justify how this idea differs from existing ones, highlighting its unique aspects. In <JSON>, provide the new idea with the following fields and provide as many details as possible: - "Idea": A detailed description of the idea, outlining its significance and potential impact. - "Title": A title for the idea, will be used for the paper writing.

[12] https://github.com/NoviScl/AI-Researcher

-"Experiment": An outline of the implementation process. Describe your high-level design plan, including necessary design steps and the ideal outcomes of the experiments.

- "Excitement": {excitement scoring rules} - "Excitement Rationale": {prompt for rationale}

- "Feasibility": {feasibility scoring rules} - "Feasibility Rationale": {prompt for rationale}

- "Novelty": {novelty scoring rules} - "Novelty Rationale": {prompt for rationale}

Be cautious and realistic on your ratings. This JSON will be automatically parsed, so ensure the format is precise, and the content should be longer than 600 words. You only need to output one idea.

Table 7 (continued). An example for research idea generation.

Title: "Multimodal Few-Shot Learning for Cognitive Bias-Based Detection of Machine-Generated Text",
Idea: "This idea proposes a multimodal, few-shot learning framework for detecting machine-generated text by leveraging cognitive biases and contextual cues across text, speech, and visual modalities. Unlike existing methods that focus solely on linguistic features, this approach integrates psychological patterns (e.g., confirmation bias, anchoring) with multimodal data to create a more robust and adaptive detection system. The framework involves: (1) collecting a diverse dataset of human and machine-generated content annotated for cognitive biases and contextual features, (2) developing a multimodal deep learning model that combines text, speech, and visual embeddings, and (3) employing few-shot learning techniques to enable the model to generalize across new domains and languages. The system will be evaluated in real-time scenarios, such as social media moderation and customer service chatbots, to assess its effectiveness in distinguishing machine-generated content from human-authored text. This approach has the potential to significantly improve the accuracy and adaptability of machine-generated text detection, particularly in dynamic and multilingual environments."
Experiment : "1. Dataset Collection: Curate a multimodal dataset containing text, speech, and visual data from both human and machine-generated sources. Annotate the dataset for cognitive biases and contextual features. 2. Model Development: Design a multimodal deep learning model that integrates text, speech, and visual embeddings using transformer architectures. Incorporate few-shot learning techniques to enable domain adaptation. 3. Training and Evaluation: Train the model on the annotated dataset and evaluate its performance in detecting machine-generated text across diverse scenarios, including social media moderation and customer service chatbots. 4. Real-Time Testing: Deploy the model in real-time applications to assess its effectiveness and adaptability. Measure accuracy, recall, and F1-score as key performance metrics. 5. Comparative Analysis: Compare the

proposed model with state-of-the-art methods to demonstrate its superiority in terms of accuracy and adaptability."
Excitement: 9, Excitement Rationale: 'This idea is highly exciting because it addresses a critical challenge in NLP and multimodal learning by integrating cognitive biases with advanced machine learning techniques. It has the potential to significantly improve the detection of machine-generated text in real-time applications, which is increasingly important in the era of large language models and AI-generated content. The combination of multimodal data and few-shot learning makes this approach highly adaptable and generalizable, setting it apart from existing methods.",
Feasibility: 7, Feasibility Rationale: 'The idea is feasible but requires careful planning and efficient use of resources. While the multimodal dataset collection and model development are achievable within 1-2 months, the real-time testing and deployment may require additional computational resources. The use of APIs and pre-trained models can help mitigate some of these challenges, but the integration of few-shot learning techniques may require advanced expertise."

Table 8. Prompt for comparison of research ideas.

System prompt: You are a reviewer specialized in Natural Language Processing. You are given two project summaries. One of them is accepted by a top AI conference (like ICLR or ACL) and the other one is rejected. Your task is to identify the one that has been accepted.
Now decide which one is the accepted idea. give a meta-review to each paper. ```json{ Decision: <DECISION> ReviewForPaper1: <ReviewForPaper1> ReviewForPaper2: <ReviewForPaper2> } ``` In <DECISION>, return a number 1 or 2 and nothing else In <ReviewForPaper1>, write the review for paper 1,reasons for acceptance or rejection. In <ReviewForPaper2>, write the review for paper 2,reasons for acceptance or rejection. This JSON will be automatically parsed, so ensure the format is precise.

Table 9. Prompt for research idea abstract generation.

System prompt: You are now a senior scholar in the field of artificial intelligence..
You are an ambitious scientist who will generate a summary based on given research idea and experimental steps. Requirements: The content of the abstract should cover: research questions and objectives, research methods, expected research results, and conclusions. Do not exceed 300 words.

Here is the research idea: '"{idea}"'
Here is the experimental steps: '"{experiment}"'
"Please respond in the following format:
Thought: <THOUGHT>
Abstract: ```json<JSON>```
In <THOUGHT>, please briefly describe your thinking.
In <JSON>, provide the abstract with the following fields:
- "Title": A title for the abstract.
- "Abstract": abstract.
Be cautious and realistic on your ratings. This JSON will be automatically parsed, so ensure the format is precise.

Table 10. An example for research idea abstract generation.

Thought: The research idea focuses on developing a robust framework for detecting AI-generated content across multiple modalities by leveraging cognitive biases and advanced fusion techniques. The experimental steps outline a systematic approach to dataset collection, model development, fusion mechanism implementation, domain adaptation, evaluation, and deployment. The abstract should succinctly capture the research objectives, methods, expected results, and conclusions while maintaining clarity and precision.

Abstract:
{
"Title": "Contrastive Meta-Style Adversarial Fusion Network for Robust Detection of AI-Generated Content",
"Abstract": "This research proposes the Contrastive Meta-Style Adversarial Fusion Network (CMSAF-Net), a novel framework designed to detect AI-generated content across text, audio, and visual modalities by leveraging cognitive biases. The framework integrates modality-specific bias detection models, a contrastive hierarchical fusion mechanism, and a meta-style adversarial domain adaptation module. The methodology involves dataset collection, training of modality-specific models using advanced NLP, speech analysis, and computer vision techniques, and the development of a hierarchical fusion algorithm enhanced with contrastive learning. Additionally, meta-style adversarial training is employed to ensure generalization across diverse datasets. The framework is evaluated on a held-out dataset, with performance metrics including accuracy, precision, recall, and F1-score. Expected results indicate that CMSAF-Net will outperform existing unimodal and multimodal detection methods, providing a robust and generalizable solution for detecting synthetic media. The research concludes with the deployment of a user-friendly interface, aiming to mitigate the challenges posed by AI-generated content in misinformation campaigns."
}

https://doi.org/10.51408/issi2025_016

Evaluating the Obsolescence Patterns in Early and Non-Early Publications: The Role of Open Access and Document Type

Huei-Ru Dong[1], Mu-Hsuan Huang[2]

[1]*141646@mail.fju.edu.tw*
Fu Jen Catholic University, Dept of Library and Information Science, New Taipei (Taiwan)

[2]*mhhuang@ntu.edu.tw*
National Taiwan University, Dept of Library and Information Science, Taipei City (Taiwan)
National Taiwan University, Center for Research in Econometric Theory and Applications (CRETA), Taipei City (Taiwan)

Abstract

This study evaluates the obsolescence patterns in early and non-early publications within the field of library and information science, focusing on the role of open access (OA) and document type. The findings reveal distinct patterns in the dissemination and longevity of research. Early publications exhibit a higher preference for OA, reflecting a trend towards rapid and accessible dissemination of research findings. The citation half-life analysis indicates that non-OA early publications tend to have a longer citation lifespan compared to OA early publications, while OA non-early publications demonstrate a more extended impact than their non-OA counterparts. Document type analysis shows that 'Article' and 'Review' papers are predominantly early publications, suggesting these formats are prioritized for early release due to their comprehensive and impactful nature. The study underscores the evolving landscape of scholarly communication, highlighting the increasing adoption of OA and its implications for research visibility and longevity. Future research should expand to other disciplines, extend the temporal scope, and incorporate a broader range of impact metrics, such as social media mentions and altmetrics. Additionally, differentiating between types of OA (e.g., gold, green, hybrid) could provide more nuanced insights into their respective impacts on citation and dissemination patterns. This research emphasizes the importance of OA in enhancing the accessibility and impact of scholarly work, while also identifying areas for further exploration to better understand the dynamics of early and non-early publications.

Introduction

The lifecycle of information involves its creation, documentation, dissemination, usage, and eventual decline in use. From an information science perspective, the value of documented information diminishes over time, a process known as literature obsolescence (Gosnell, 1943). Literature obsolescence is primarily defined as the decrease in the usage of papers, typically assessed through citation or being cited. However, obsolescence is a dynamic process and challenging to capture. Therefore, the measurement of literature obsolescence often borrows the concept of half-life from physics, defining the half-life of a paper as the time required for its usage to reach half of its total usage.

Generally, the measurement of literature obsolescence can be conducted using two methods: synchronous and diachronous. The synchronous method explains the phenomenon of obsolescence at a specific point in time, focusing on the age distribution of cited information. This can be obtained by calculating the median age of cited references. The diachronous method observes a specific group of papers and

measures the time required for their citations to reach half of their total citations after publication. This can be obtained through citation data of the papers.

With the digital age transforming the academic publishing process, many journal articles are often disclosed as "early access" before their official publication. This necessitates considering whether the "early publication date" significantly impacts the assessment of literature obsolescence. Since December 2017, the WoS database has provided early publication date information and bibliographic fields, including Early Access Date (EA) and Early Access Year (EY). Once a paper is officially published, the EA and EY information is retained, and the official publication year and date are added (Clarivate, 2020b). However, the EY bibliographic field is no longer provided.

Many studies have explored the differences in various publication dates, such as early access date, official publication date, and database indexing date (Alves-Silva et al., 2016; Das & Das, 2006; Maflahi & Thelwall, 2018). There have also been comparisons of publication delays and early access (Al & Soydal, 2017; Gonzalez-Betancor & Dorta-Gonzalez, 2019; Heneberg, 2013; Hu et al., 2018; Kousha et al., 2018; Liu et al., 2020).

Dong et al. (2024) investigated the differences in literature obsolescence assessment caused by time lag, using the field of library and information science as an example. They analysed trends, citation impact, timeliness, and time lag of early publications and non-early publications. The results showed that the number of early publications has steadily increased each year, while non-early publications have shown a declining trend, although their numbers are still higher than early publications. Despite being fewer in number, early publications have a higher citation impact, with significantly higher average citation counts than non-early publications. Additionally, the citation half-life of early publications is about seven years or longer, indicating a longer citation lifespan, whereas the citation half-life of non-early publications is one to three years. Finally, the study found that more than half of early publications are officially published within three months, and about a quarter are exposed nine months in advance. This study highlights the importance of early publication in enhancing the impact and visibility of academic research.

Open Access refers to the authorized provision of free access to the full text of academic papers, which has become a significant trend in academic publishing in the internet age. It helps facilitate academic dissemination and promotes academic freedom. Because Open Access (OA) journal articles have lower access barriers, they tend to have higher visibility and citation opportunities compared to subscription-based journal articles. Consequently, many studies have explored and compared the citation impact advantage of OA versus non-OA journal articles. Most research suggests that OA journal articles indeed have a citation impact advantage (Eysenbach, 2006; Gargouri et al., 2010; Harnad & Brody, 2004; Norris et al., 2008), and they are cited more quickly (Atchison & Bull, 2015). However, some researchers argue that OA journals do not have a citation impact advantage (Davis et al., 2008; Davis & Walters, 2011; Dorta-González et al., 2017; Moed, 2007; Sotudeh, 2020). Additionally, some scholars believe that this discrepancy is due to differences across disciplines (Hubbard, 2017).

Document type refers to the format in which a paper is published. According to the SSCI database, there are 13 types of documents, roughly ranked by quantity as follows: Article, Meeting Abstract, Review, Editorial Material, Proceedings Paper, Book Review, Letter, Correction, Book Chapter, Biographical-Item, Retracted Publication, News Item, and Retraction (Clarivate, 2020a).

Different types of papers can lead to variations in citation counts. For example, theoretical or empirical research papers tend to be cited less frequently than methodological papers. The differences in document types also significantly impact citation counts, making it necessary to consider document types when conducting citation analysis (Peritz, 1983). Hamilton (1991) also pointed out that papers that are not cited may be obituaries, short notes, reviews, communications, or other types of documents. These types of papers often do not contain rigorous experimental or survey results, which explains why they are cited less frequently.

Based on the study by Dong et al. (2024) that investigates the obsolescence and time lag of early publications and non-early publications, this research aims to further explore the obsolescence patterns of EA and non-EA papers from the perspectives of Open Access and document types. The research objectives are as follows:

1. Investigate the Impact of Open Access on the Citation Half-Life of Early and Non-Early Publications: Examine how open access status affects the citation half-life of early publications and non-early publications.
2. Analyze the Effect of Various Document Types on the Citation Half-Life of Early and Non-Early Publications: Explore how different document types influence the citation half-life of early publications and non-early publications.

Methodology

This study investigates the obsolescence patterns in Early Publications and Non-Early Publications. It examines whether there are differences in the aging of literature between Early Publications and Non-Early Publications based on two attributes: whether the publication is Open Access and the document type of the publication. The following sections will detail the research methods, data acquisition, and data processing of this study.

Data collection

This study retrieved and downloaded bibliographic data from the SSCI database nn June 20, 2023. The search query was "WC=Information Science & Library Science," with the publication date limited to between January 1, 2013, and December 31, 2022. SSCI only started providing early publication dates for papers in 2019, and the early publication date data in the downloaded bibliographies is not comprehensive.

Therefore, this study used Python programming language to extract the online publication date of each article in journals within the field of Library and Information Science from the journal's website. The online publication dates obtained from the websites were combined with the early publication dates in the bibliographies to

form the early publication date data for this study. The research samples were divided into two categories: early publication papers and non-early publication papers.

(1) **Early publication**: Refers to bibliographies in the field of Library and Information Science with early publication dates. The early publication dates mainly come from journal websites, with some from WoS bibliographic data. There is a total of 33,748 early publications.
(2) **Non-early publication**: Refers to bibliographies in the field of Library and Information Science without early publication dates. This study includes a total of 63,450 non-early publication papers.

Literature Obsolescence

This study uses the citation half-life to measure literature obsolescence. The citation half-life is defined as the median age difference between the publication year of a set of papers and the publication year of their cited references. A smaller value indicates a faster rate of literature obsolescence, meaning the knowledge in that set of papers is updated more quickly. Conversely, a larger value indicates a slower rate of literature obsolescence, meaning the knowledge in that set of papers is updated more slowly.

The calculation method involves first sorting the cited references of each paper by publication year. Then, the publication year of the median cited reference is obtained. The time difference between the official publication year and the early publication year of each paper is calculated. The average value of the citation half-life is then computed for groups of papers based on whether they are Open Access (OA) or by document type.

Open Access and Non-Open Access

Whether a journal is OA is determined by the OA field in the SSCI bibliographic data. If the OA field has a value, the paper is considered Open Access; if the OA field is empty, the paper is considered Non-Open Access.

Document Type

The document type is based on the DT field in the SSCI bibliographic data. There are 1,752 bibliographic records that include more than one document type. Among them, there are 1,303 early publications and 449 non-early publications.

Results

Distribution and Ratio of Early Publications and Non-Early Publications Between Open Access and Non-Open Access

Table 1 compares the number of early publications and non-early publications between open access and non-open access in the field of library and information science from 2013 to 2022. In early publications, there are 12,119 open access papers, accounting for 35.91% of early publications. However, in non-early publications, there are only 7,434 open access papers, accounting for just 11.72% of non-early publications. Early publications not only have more open access papers, but the

proportion of open access papers is also higher. Regardless of whether they are early publications or non-early publications, the number and proportion of non-open access papers are higher. Particularly in non-early publications, the proportion of non-open access papers is as high as 88.28%.

Looking at the annual trend, in early publications, the number of open access papers shows a gradual increase, with a slight increase in proportion. The number of non-open access papers also shows a gradual increase, but the proportion slightly decreases. As for non-early publications, the number and proportion of open access papers remain relatively stable, while the number and proportion of non-open access papers show a slight decrease. Overall, whether open access or non-open access, the number and proportion of early publications are gradually increasing.

Table 1. The Number and Ratio of Early Publications and Non-Early Publications Across Open Access (OA) and Non-Open Access (non-OA) Papers.

Publication	*Early publications*		*Non-early publications*	
Year	*OA (%)*	*non-OA (%)*	*OA (%)*	*non-OA (%)*
2013	794 (30.54%)	1,806 (69.46%)	713 (10.01%)	6,411 (89.99%)
2014	915 (31.31%)	2,007 (68.69%)	651 (9.01%)	6,571 (90.99%)
2015	932 (31.49%)	2,028 (68.51%)	756 (10.00%)	6,802 (90.00%)
2016	1,183 (38.38%)	1,899 (61.62%)	813 (11.00%)	6,576 (89.00%)
2017	1,194 (38.04%)	1,945 (61.96%)	745 (10.71%)	6,214 (89.29%)
2018	1,204 (38.37%)	1,934 (61.63%)	845 (13.50%)	5,416 (86.50%)
2019	1,128 (36.39%)	1,972 (63.61%)	744 (14.57%)	4,362 (85.43%)
2020	1,462 (39.40%)	2,249 (60.60%)	762 (15.64%)	4,111 (84.36%)
2021	1,769 (40.20%)	2,632 (59.80%)	710 (14.43%)	4,211 (85.57%)
2022	1,538 (32.76%)	3,157 (67.24%)	698 (11.56%)	5,340 (88.44%)
Total	12,119 (35.91%)	21,629 (64.09%)	7,437 (11.72%)	56,014 (88.28%)

Citation Half-Life Comparison Across Open Access and Non-Open Access

Table 2 compares the citation half-life of early publications and non-early publications between open access and non-open access in the field of library and information science from 2013 to 2022. Over the 10-year citation half-life, the citation half-life of non-open access early publications (7.73) is higher than that of open access early publications (6.40). However, for non-early publications, the citation half-life of open access papers (7.07) is higher than that of non-open access papers (1.38).

Looking at the annual trend, the 10-year citation half-life for early publications, whether open access or non-open access, remains relatively stable. For non-early publications, the citation half-life of open access papers shows a slight upward trend each year, while the citation half-life of non-open access papers shows a slight downward trend each year. In summary, early publications have a higher citation half-life, with non-open access early publications having a slightly higher citation half-life than open access early publications.

Table 2. The citation half-life of Early Publications and Non-Early Publications for open access and non-open access.

Publication Year	*Early publications*		*Non-early publications*	
	OA	*non-OA*	*OA*	*non-OA*
2013	5.85	7.43	6.14	1.47
2014	6.19	7.70	6.26	1.22
2015	6.38	7.88	7.46	1.11
2016	6.41	8.14	6.52	1.35
2017	6.76	7.97	6.64	1.29
2018	6.39	7.80	7.35	1.27
2019	6.79	7.79	7.59	2.04
2020	6.52	7.71	7.82	1.57
2021	6.16	7.61	7.50	1.80
2022	6.46	7.56	7.91	1.15
Average	6.40	7.73	7.07	1.38

Distribution and Ratio of Early Publications and Non-Early Publications Across Various Document Types

Table 3 presents the number and ratio of early publications and non-early publications of various document types in the field of library and information science. The table only includes document types with a total paper count exceeding 500 over ten years. It reveals that the document type 'Article' has the highest number of papers in the field, totaling 43,416. The second most abundant document type is 'Book Review' with a total of 43,012 papers, followed by 'Editorial Material' with 6,424 papers. The remaining document types have fewer than 2,000 papers each.

In terms of early publications, the number of 'Article' papers remains the highest, with the ratio of early publications reaching 66.74%, indicating that 'Article' papers tend to be published early. Although the number of early publications for 'Review' is not the highest, the ratio is the highest at 75.18%, with as many as three-quarters of 'Review' papers being published early. For 'Letter' and 'Proceedings Paper', although the numbers are not large, more than 30% of the papers are early publications. On the other hand, the ratio of early publications for 'Book Review' and 'News Item' document types is quite low, especially for 'News Item', where all papers are non-early publications with no early publications at all.

Table 3. The Number and Ratio of Early Publications and Non-Early Publications Across Various Document Types.

Document Types	*Total*	*Early publications (%)*	*Non-early publications %)*
Article	43,416	28,974 66.74%)	14,442 33.26%)
Book Review	43,012	1,448 3.37%)	41,564 96.63%)
Editorial Material	6,424	1,505 23.43%)	4,919 76.57%)
Review	1,640	1,233 75.18%)	407 24.82%)
News Item	1,088	0 0.00%)	1,088 100.00%)
Letter	832	302 36.30%)	530 63.70%)
Proceedings Paper	733	284 38.74%)	449 61.26%)

Citation Half-Life Comparison Across Various Document Types

Table 4 compares the citation half-life of total papers, early publications, and non-early publications of various document types in the field of library and information science from 2013 to 2022. For total papers, 'Proceedings Paper,' 'Article,' and 'Review' have relatively high citation half-lives, all above 7. The citation half-life of 'Book Review' is only 0.37, significantly lower than other document types. This is because 'Book Review' mainly cites newly published books.

Comparing early publications and non-early publications, except for 'Letter' and 'Book Review,' early publications generally have a lower citation half-life. Most document types have a higher citation half-life for non-early publications. Particularly for 'Proceedings Paper,' 'Article,' and 'Review,' the citation half-life of non-early publications is above 8. In conclusion, non-early publications generally exhibit a higher citation half-life compared to early publications across most document types.

Table 4. The citation half-life of Early Publications and Non-Early Publications for various document types.

Document Types	*Total publications*	*Early publications*	*Non-early publications*
Proceedings Paper	7.97	7.03	**8.54**
Article	7.81	7.63	**8.39**
Review	7.19	7.02	**8.21**
News Item	6.81	0.00	**6.81**
Editorial Material	5.17	4.45	**6.12**
Letter	4.68	**5.72**	1.41
Book Review	0.37	**2.68**	0.31

Conclusion and discussion

This study investigates the obsolescence patterns in Early Publications and Non-Early Publications based on Open Access and the document type of the publication. The findings reveal several key trends and patterns. Early publications have a significantly higher proportion of OA papers compared to non-early publications, indicating a preference for OA in the early dissemination of research findings. Both early and non-early publications show an increasing trend in the number of OA papers over the years, although the proportion of OA papers in non-early publications remains relatively stable. The citation half-life of non-OA early publications is higher than that of OA early publications. Conversely, for non-early publications, OA papers have a higher citation half-life compared to non-OA papers. The 'Article' document type has the highest number of papers, with a significant proportion being early publications. 'Review' papers have the highest ratio of early publications, while 'Book Review' and 'News Item' have the lowest ratios of early publications.

The findings of this study highlight the evolving landscape of scholarly communication in the field of library and information science. The higher proportion

of OA in early publications suggests that researchers are increasingly opting for OA to ensure rapid dissemination and wider accessibility of their work. This trend aligns with the broader movement towards open science and the push for greater transparency and accessibility in research. The increasing trend in OA papers, particularly in early publications, reflects the growing acceptance and adoption of OA publishing models. This shift is likely driven by several factors, including the increasing availability of OA journals, mandates from funding agencies, and the perceived benefits of OA in terms of visibility and impact.

The citation half-life analysis provides insights into the longevity and impact of OA and non-OA publications. The higher citation half-life of non-OA early publications suggests that these papers continue to be cited over a longer period, possibly due to their perceived quality or the prestige of the journals in which they are published. However, the higher citation half-life of OA non-early publications indicates that OA papers in this category also have a lasting impact, likely due to their accessibility and visibility. The analysis of document types reveals that 'Article' and 'Review' papers are the most common and have the highest ratios of early publications. This suggests that these document types are prioritized for early dissemination, possibly due to their comprehensive nature and the critical role they play in advancing knowledge in the field. On the other hand, 'Book Review' and 'News Item' document types have lower ratios of early publications, indicating that these types of documents are less likely to be published early.

Despite the comprehensive nature of this study, several limitations should be acknowledged. The study is limited to the field of library and information science and may not be generalizable to other disciplines. Future research could expand the scope to include other fields to provide a more holistic view of OA and non-OA publishing trends. The analysis covers a ten-year period from 2013 to 2022. While this provides a substantial dataset, extending the temporal coverage could capture longer-term trends and changes in publishing practices. The study relies on citation half-life as a measure of impact. While this is a useful metric, it does not capture other dimensions of impact, such as social media mentions, downloads, or altmetrics. Future studies could incorporate a broader range of impact metrics to provide a more comprehensive assessment. The study categorizes papers as OA or non-OA based on their availability. However, there are different types of OA (e.g., gold, green, hybrid) that may have varying impacts on citation and dissemination. Future research could differentiate between these types to provide more nuanced insights.

In conclusion, this study sheds light on the dynamics of OA and non-OA publishing in the field of library and information science, highlighting key trends, patterns, and areas for future research. The findings underscore the importance of OA in the early dissemination of research and its potential impact on the visibility and longevity of scholarly work.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 113L900202) which is under the Featured Areas Research Center Program by Higher Education Sprout Project of Ministry of

Education (MOE) in Taiwan, the Universities and Colleges Humanities and Social Sciences Benchmarking Project (Grant no. 113L9A001).

References

Al, U., & Soydal, I. (2017). Publication lag and early view effects in information science journals. *Aslib Journal of Information Management*, *69*(2), 118–130. https://doi.org/10.1108/AJIM-12-2016-0200

Alves-Silva, E., Porto, A. C. F., Firmino, C., Silva, H. V., Becker, I., Resende, L., Borges, L., Pfeffer, L., Silvano, M., Galdiano, M. S., Silvestrini, R., & Moura, R. (2016). Are the impact factor and other variables related to publishing time in ecology journals? *Scientometrics*, *108*(3), 1445–1453. https://doi.org/10.1007/s11192-016-2040-0

Atchison, A., & Bull, J. (2015). Will Open Access Get Me Cited? An Analysis of the Efficacy of Open Access Publishing in Political Science. *PS: Political Science & Politics*, *48*(1), 129–137. https://doi.org/10.1017/S1049096514001668

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open Access to the Scientific Journal Literature: Situation 2009. *PLOS ONE*, *5*(6), e11273. https://doi.org/10.1371/journal.pone.0011273

Clarivate. (2020a). *Searching the Document Type Field*. Web of Science Core Collection Help. http://images.webofknowledge.com/WOKRS535R102/help/WOS/hs_document_type.html

Clarivate. (2020b). *Web of Science Core Collection: Early Access articles*. WOS Training. https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Early-Access-articles?language=en_US

Das, A., & Das, P. (2006). Delay between online and offline issue of journals: A critical analysis. *Library & Information Science Research*, *28*(3), 453–459. https://doi.org/10.1016/j.lisr.2006.03.019

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: Randomised controlled trial. *BMJ*, *337*, a568. https://doi.org/10.1136/bmj.a568

Davis, P. M., & Walters, W. H. (2011). The impact of free access to the scientific literature: A review of recent research. *Journal of the Medical Library Association : JMLA*, *99*(3), 208–217. https://doi.org/10.3163/1536-5050.99.3.008

Dong, H.-R., Huang, M.-H., & Lo, S.-C. (2024). The Obsolescence and Time Lag between Early Publications and Non-Early Publications. *Proceedings of the Association for Information Science and Technology*, *61*(1), 896–898. https://doi.org/10.1002/pra2.1132

Dorta-González, P., González-Betancor, S. M., & Dorta-González, M. I. (2017). Reconsidering the gold open access citation advantage postulate in a multidisciplinary context: An analysis of the subject categories in the Web of Science database 2009–2014. *Scientometrics*, *112*(2), 877–901. https://doi.org/10.1007/s11192-017-2422-y

Eysenbach, G. (2006). Citation Advantage of Open Access Articles. *PLOS Biology*, *4*(5), e157. https://doi.org/10.1371/journal.pbio.0040157

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLOS ONE*, *5*(10), e13636. https://doi.org/10.1371/journal.pone.0013636

Gonzalez-Betancor, S. M., & Dorta-Gonzalez, P. (2019). Publication modalities "article in press" and "open access" in relation to journal average citation. *Scientometrics*, *120*(3), 1209–1223. https://doi.org/10.1007/s11192-019-03156-2

Gosnell, C. F. (1943). *The Rate of Obsolescence in College Library Book Collections, as Determined by an Analysis of Three Select Lists of Books for College Libraries* [PhD Thesis]. New York University, School of Education.

Hamilton, D. P. (1991). Who's uncited now? *Science*, *251*(4989), 25–26.

Harnad, S., & Brody, T. (2004). Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine*, *10*(6), Article 6. https://eprints.soton.ac.uk/260207/

Heinzkill, R. (1980). Characteristics of References in Selected Scholarly English Literary Journals. *The Library Quarterly*, *50*(3), 352–365.

Heneberg, P. (2013). Effects of Print Publication Lag in Dual Format Journals on Scientometric Indicators. *PLOS ONE*, *8*(4), e59877. https://doi.org/10.1371/journal.pone.0059877

Hu, Z., Tian, W., Xu, S., Zhang, C., & Wang, X. (2018). Four pitfalls in normalizing citation indicators: An investigation of ESI's selection of highly cited papers. *Journal of Informetrics*, *12*(4), 1133–1145. https://doi.org/10.1016/j.joi.2018.09.006

Hubbard, D. E. (2017). Open access citation advantage? A local study at a large research university. *Proceedings of the Association for Information Science and Technology*, *54*(1), 712–713. https://doi.org/10.1002/pra2.2017.14505401126

Kousha, K., Thelwall, M., & Abdoli, M. (2018). Can Microsoft Academic assess the early citation impact of in-press articles? A multi-discipline exploratory analysis. *Journal of Informetrics*, *12*(1), 287–298. https://doi.org/10.1016/j.joi.2018.01.009

Liu, J., Grubler, A., Ma, T., & Kogler, D. F. (2020). Identifying the technological knowledge depreciation rate using patent citation data: A case study of the solar photovoltaic industry. *Scientometrics*. https://doi.org/10.1007/s11192-020-03740-x

Maflahi, N., & Thelwall, M. (2018). How quickly do publications get read? The evolution of mendeley reader counts for new articles. *Journal of the Association for Information Science and Technology*, *69*(1), 158–167. https://doi.org/10.1002/asi.23909

McGillivray, B., & Astell, M. (2019). The relationship between usage and citations in an open access mega-journal. *Scientometrics*, *121*(2), 817–838. https://doi.org/10.1007/s11192-019-03228-3

Moed, H. F. (2007). The effect of "open access" on citation impact: An analysis of ArXiv's condensed matter section. *Journal of the American Society for Information Science and Technology*, *58*(13), 2047–2054. https://doi.org/10.1002/asi.20663

Norris, M., Oppenheim, C., & Rowland, F. (2008). The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology*, *59*(12), 1963–1972. https://doi.org/10.1002/asi.20898

Peritz, B. C. (1983). Are methodological papers more cited than theoretical or empirical ones? The case of sociology. *Scientometrics*, *5*(4), 211–218.

Pritchard, A. (1969). Statistical bibliography or bibliometrics. *Journal of Documentation*, *25*(4), 348–349.

Sotudeh, H. (2020). Does open access citation advantage depend on paper topics? *Journal of Information Science*, *46*(5), 696–709. https://doi.org/10.1177/0165551519865489

https://doi.org/10.51408/issi2025_045

Evaluating the Scholarly Contributions of a Journal by Measuring the Discrepancy in Information Entropy Values Between Factual and Counterfactual Knowledge Systems in the Absence of the Journal

Zheng Ma[1], Liao Yu[2], Zhenglu Yu[3], Hongmei Guo[4], Ming Cheng[5]

[1]*mazheng@mail.las.ac.cn*
National Science Library, Chinese Academy of Science, Beijing, No. 33 Beisihuan West Road (P. R. China)

[2]*liaoyu@mail.las.ac.cn*
National Science Library, Chinese Academy of Science, Beijing, No. 33 Beisihuan West Road (P. R. China)

[3]*luluyu@istic.ac.cn*
Institute of Scientific and Technical Information of China, Beijing, No.15 Fuxing Road (P. R. China)

[4]*guohm@istic.ac.cn*
Institute of Scientific and Technical Information of China, Beijing, No.15 Fuxing Road (P. R. China)

[5]*mingcheng0224@163.com*
Wuhan University, Wuhan, No. 299 Bayi Road (P. R. China)

Abstract

This study proposes a novel evaluation concept and method: assessing the value of academic journals by measuring their contributions to the knowledge system. It aims to address the limitations of traditional peer review methods and quantitative approaches based on bibliometrics and altmetrics in the practical evaluation of academic journals. The study hypothesizes that academic journals play a crucial role in the knowledge system by providing valuable information through the publication of research papers, thereby reducing uncertainty within the system. As the knowledge system evolves from disorder to order, its information entropy value tends to decrease, and the academic contributions of journals can be characterized by the negentropy derived from these publications. The study employs the concept of counterfactual research to calculate the information entropy of both the factual knowledge system and the counterfactual knowledge system in the absence of the evaluated journals. The difference in information entropy values indicates the negative entropy contributed by the evaluated journals to the knowledge system. Through empirical data, this study demonstrates that this innovative method can effectively reflect the value of journals based on their actual contributions, and it has the potential to complement traditional evaluations of journal value based on impact after further refinement. The empirical data also reveal that, in general, a small number of journals within each discipline make significant contributions to the knowledge system, while the majority of journals contribute little or nothing. This finding aligns with the nucleus zone of periodicals described by Bradford's Law.

Introduction

Academic journals serve as the primary platform for documenting innovative achievements and scientific research findings. Consequently, in the realm of scientific governance and academic communication, it is essential to design rational, scientific, and accurate evaluation methods for academic journals. Initially, the evaluation of academic journals primarily relied on peer review (Baldwin 2017). However, contemporary evaluation methods increasingly emphasize quantitative approaches, which can be broadly categorized into two types: traditional bibliometrics and forward altmetrics (Karanatsiou 2017).

It is generally believed that at the beginning of the 1900s, the development of industrial technology and the rapid emergence of academic dissemination activities led to a significant increase in the volume of academic literature and the variety of journals (Huang 2021). The economics need for evaluation to identify important journals became a priority (Lewis 1989), prompting the exploration of journal evaluation methodologies, which began in Europe. Bradford (1934) summarized the law of scattering, discovering that each subject area has a nucleus zone of periodicals that publish the majority of articles within that field. The theory of academic journal evaluation also originated from Bradford's law regarding the stratification of academic journals. In the 1950s, Garfield (1955) pioneered the establishment of a citation analysis system, gradually developing a series of citation databases and expanding their practical applications. This work led to the formulation of a comprehensive analysis system and methodology, which has had a significant impact on the field (Vinkler 2009). However, traditional scientometrics indicators based on citation analysis also present notable challenges. For instance, citation analysis often requires a lengthy post-publication period, typically taking several years to adequately assess the academic influence of journals (Feng 2023). Additionally, the evaluation data sources for traditional scientometrics indicators primarily focus on quantitative metrics, such as the number of articles or citations, while neglecting the roles and impacts that evaluated journals have in areas such as academic exchange, industrial development, and disciplinary advancement (Wang 2011). The fundamental assumption of citation analysis is that citations reflect the positive impact of academic contributions (Narin 1990); however, in practice, the motivations for citing a particular paper are more varied, and citing a work does not necessarily indicate that the citer endorses it (Dorta-Gonzalez 2013).

Priem and Taraborelli (2010) co-authored a paper titled *Altmetrics: A Manifesto*, which introduced the concept of altmetrics. The scientific and effective application of alternative metrics facilitates a more comprehensive evaluation of impact (Shuai 2012). In terms of evaluation orientation, the use of alternative metrics will foster more vibrant and efficient scientific exchanges on the Internet (Eysenbach 2011) and can lead to the development of new methods, tools, and mechanisms to enhance and optimize existing information organization and discovery processes (Priem 2012). However, a significant challenge in applying alternative metrics to the evaluation of scientific and technical journals is minimizing human interference with the metrics (Bornmann 2014). Related concerns also include the rigor and consistency of data used in alternative measures (Cronin 2014). Another important issue is how to ensure

that widely dispersed and dynamic data sources are reliable (Maflahi 2016) and that the results of their statistical analyses are reproducible (Thelwall 2013).

Despite numerous explorations, an effective solution to the aforementioned limitations within the current evaluation model of academic journals remains elusive. Therefore, Ma (2022) believes that future approaches to evaluating academic journals will transcend from the traditional framework of statistical analysis focused on the journals' inherent attributes and external connections into a systematic perspective that quantitatively assesses the actual contributions of the publishing and dissemination behaviors of the evaluated journals to the evolution of the knowledge systems in which they operate.

This study posits that one of the primary roles of academic journals is to mitigate uncertainty in scientific understanding. The process of reducing uncertainties in scientific knowledge corresponds to a decrease in entropy within the knowledge system (Shannon 1963). The fundamental purpose of the academic publishing process is to enhance individuals' awareness of scientific issues and principles through the dissemination and promotion of scientific discoveries and technological innovations. Utilizing quantitative methods, this study measures the changes in information entropy within the knowledge system before and after academic publishing and develops an evaluation method to assess the contributions of scientific and technical academic journals. The degree of negentropy that an academic journal introduces to the information entropy of the knowledge system reflects its contribution to the advancement of the discipline. In this study, we propose a solution to measuring the utility of information by examining the discrepancies in information entropy values between factual and counterfactual knowledge systems, based on the concept of counterfactual thinking (Kahneman 1982).

Concepts defined in this study

Knowledge systems

The knowledge system, formed by journal articles, is defined in this study as a framework that consolidates explicit human perceptions of the objective world within specific boundaries. This system is based on various knowledge carriers and encompasses both similar and differing research perspectives. Over time, this system experiences changes, additions, and the disappearance of certain perspectives.

Uncertainty in knowledge systems

In a knowledge system, the variations in the composition of individual research perspectives are regarded as the inherent uncertainty within the system. This uncertainty can be categorized into two types: static uncertainty and dynamic uncertainty.

(1) Distribution uncertainty (static level)

Distribution state uncertainty primarily reflects, at a static level, whether the distribution of absolute indicators within a knowledge system's conclusions about

academic content and the convergence of research concerns is significantly centralized or decentralized. Over time, in a given system, the more consistent and concentrated the judgments regarding knowledge viewpoints, research hotspots, and mainstream development directions are, the clearer and more coherent the knowledge system's understanding of academic issues becomes. This indicates a more complete and accurate human comprehension of the objective world. Conversely, if the exploration and understanding of knowledge within a system are more diverse, and the probabilities of different directions and conclusions are relatively similar, it suggests that human understanding of the relevant issues remains uncertain, lacking clarity and consistency.

The formulation of static uncertainty within the knowledge system, constructed from academic journal articles, can be further decomposed into three subsystems.

A1. Scalability: This term refers to the capacity of academic journals to effectively disseminate literature. As publishing and communication platforms, academic journals should aim to publish a significant number of papers that showcase the results of scientific discoveries and technological innovations, all while upholding high standards of quality and efficiency.

A2. Wideness: This term refers to the ability of an academic journal to broaden its influence. The content published and disseminated by academic journals consists of scientific research papers, which require a substantial readership to effectively share results and promote active academic communication.

A3. Sustainability: This term primarily refers to the quantity and proportion of papers funded by financial support, serving as an indicator of the alignment between journal publications and scientific and technological investments.

(2) Relation uncertainty (dynamic level)

Relation state uncertainty primarily reflects a dynamic knowledge system concerning the structure of nodes related to academic knowledge, the interactions between these nodes, and whether the relationships among different types of nodes indicate a centralized or decentralized state. Within this system, various node levels (e.g., authors, journals, keywords, individual papers) form a network of connections that represent knowledge. The relative centralization of the entire knowledge system can be inferred from the connections between knowledge nodes and their related nodes as expressed by this network. Absolute centralization implies that when people's judgments regarding knowledge perspectives, research hotspots, and mainstream development directions are highly consistent and concentrated, their understanding of academic issues within this knowledge system becomes clearer and more uniform. In other words, human comprehension of the objective world tends to be completer and more accurate. Conversely, if a system's exploration and understanding of knowledge exhibit greater diversity, and the likelihood of different directions and conclusions is relatively similar, it indicates that human understanding of the issue remains uncertain, lacking clarity and consistency.

The systematic uncertainty associated with the relatively centralized knowledge system constructed from academic journal papers can be further decomposed into four subsystems.

B1. Openness: This term refers to the transparency and accessibility of manuscript sources for journal articles. The development of manuscript sources is a critical aspect of establishing academic journals. A diverse and ample supply of high-quality manuscript sources is essential for journals to effectively fulfill their roles. Conversely, if the range of manuscript sources is overly restricted or concentrated, it may lead to a one-sided knowledge system and can diminish the communicative vitality of academic journals.
B2. Collaboration: This term refers to the capacity of journals to publish co-authored papers, including those arising from collaborative research at both national and institutional levels. Collaborative research often yields complementary advantages and generates high-quality research outcomes. Notably, the large-scale multilateral collaborations that have surged in recent years have led to the production of papers, which frequently contain key findings that can benefit the global community.
B3. Competitiveness: The capacity of a specific journal to achieve a comparative advantage over other journals within the same discipline or genre. High-competitiveness journals typically attract high-quality research, establishing authority and influence (Ma and Pan etc. 2022). This authority and influence contribute to the formation of academic consensus, thereby reducing disagreements and uncertainties regarding certain issues.
B4. Influence: This term refers to the reference value or contentious significance of the results published in a thesis by academic journals, particularly in relation to other scholarly research activities. It is primarily measured by the number of citations. The citations of a journal article serve as a key indicator of the academic impact of the paper.

Information Entropy of Knowledge Systems

It is due to the significant systemic properties of disciplinary development and dissemination that a collection of papers published in academic journals within a specific subject area can be analyzed as a relatively independent system. In this study, the information entropy of the knowledge system is defined as follows: within the closed and isolated knowledge system formed by the research papers of the journals, the measure of uncertainty regarding the knowledge and judgment of a particular scientific problem is defined as the information entropy of this knowledge system.

Counterfactual knowledge system

Counterfactual knowledge systems are virtual constructs, in contrast to real knowledge systems. This concept assumes that the evaluated journal does not exist; that is, the journal is excluded from the real knowledge system. Consequently, the volume of its published papers, references, and citations is not factored into the statistical calculations of relevant data and indicators.

Discrepancy in Information Entropy values between Factual and Counterfactual Knowledge Systems

For the evaluated journal, there is a discrepancy in information entropy values between factual knowledge systems and counterfactual knowledge systems that do

not include the evaluated journal. The primary reason for this gap is the contribution of the evaluated journal, which reduces the information entropy of the knowledge system. In other words, the scholarly papers published by the evaluated academic journal contribute negentropy to the knowledge system.

Data

Sample

The sample of journals utilized for empirical evidence in this study comprises 3,713 scientific and technical academic journals, representing the vast majority of academic journals published in China. All of these journals were recognized by the state publishing administration of China in 2014. In this study, all data regarding journal articles and citations were downloaded from the China Journal Network (COJ) of Wanfang Data Co. (Ma 2008). The COJ includes over 8,000 journals and 43 million articles published in China, featuring high-quality full-text records that provide extensive information related to the articles. In this study, the analysis will be conducted using papers published between 2016 and 2019 as examples.

Disciplinary categories

The classification of 112 disciplinary categories is based on the *Chinese Science and Technology Journal Citation Reports (Core Edition)* (Pan and Ma 2018), *National Standard of PRC: Classification and Code of Disciplines (GB/T 13745-2009)* (State Bureau of Quality and Technical Supervision of PR China 2009), and the *Chinese Library Classification* (Editorial Committee of CLC 2010). The classification of these categories considers the affiliation of each discipline as well as the volume of publications, organized into six major parts of multidiscipline, basic research, agriculture, medicine, engineering and technology, and management.

High-frequency keywords

The set of high-frequency keywords was used as a framework for developing disciplinary options within each journal's subject area. The frequency of these keywords is derived from the CSTPCD, a WOS-like citation index for scientific and technical journals in China (Zhou 2007). The CSTPCD includes more than 2000 nucleus journals, representing approximately one-third of the total number of science and technology journals in China. Based on the CSTPCD, the high-frequency keywords that fall within the top 1% of usage frequency in each discipline are identified.

Method

As illustrated in Figure 1, this study aims to develop a quantitative model for calculating the information entropy of a knowledge system. In this context, academic journals are treated as a knowledge system, with high-frequency keywords serving as variables that represent system uncertainty. Additionally, we introduce measurable subsystem indicators. For the purpose of journal evaluation, we calculate the information entropy of both the factual knowledge system and the counterfactual

knowledge system, which assumes the absence of the evaluated journals. The difference between these two values represents the negentropy contributed by the evaluated journal to the system, reflecting its role in reducing the uncertainty of the knowledge system. This metric can be utilized to assess the academic quality and value of journals.

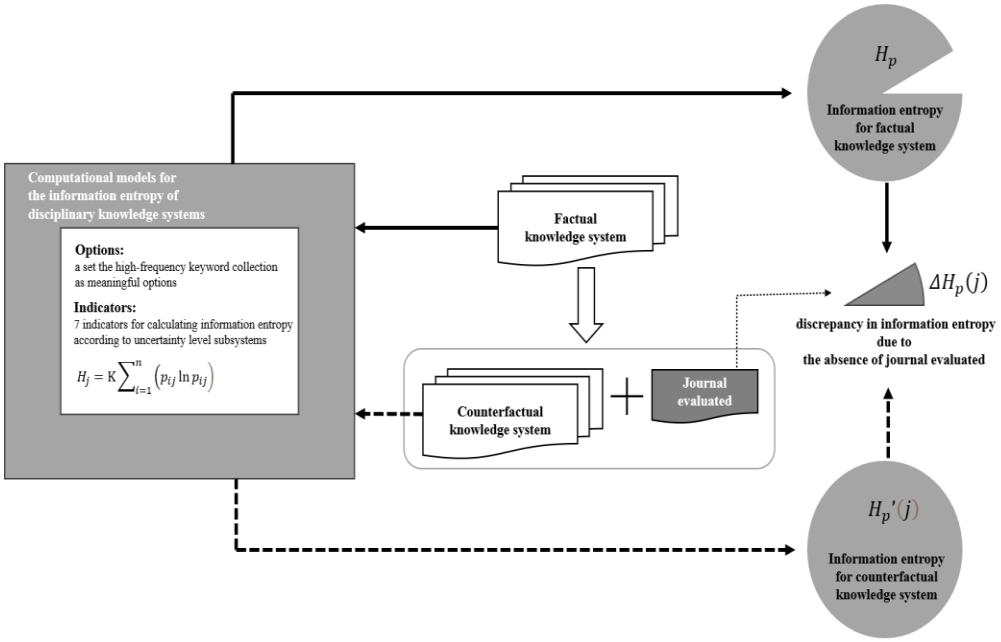


Figure 1. Research idea for this study.

To calculate the set of discipline development options

This study employs a set of high-frequency keywords for each discipline to delineate potential avenues for development within those fields. When applied to the evaluation of journals in each subject area, the model reveals varying numbers of research directions that each discipline may encompass. In other words, the number of possible options (variables) for disciplinary development differs based on the size and characteristics of each discipline.

Keywords are a set of words that express the selection, solution, technical approach, object of study, innovative ideas, application value, and other relevant aspects of a paper. According to journal publishing standards, keywords are an essential component of academic papers. They possess characteristics of standardization and universality. Typically, keywords are preferred over narrative words; that is, they should consist of semantically related and scientifically relevant terms derived from natural language vocabulary. While free words can also be utilized as keywords, it is advisable to select terms from established lexicons or widely recognized reference books and toolkits.

The keywords of a paper can reflect the direction of the chosen topic, the research methodology, or the main findings. Utilizing big data technology, the study of

keywords can facilitate an intuitive understanding of the knowledge structure and the development of the field. By analyzing the evolution of the quantitative relationships among keywords, researchers can also identify and monitor the emerging hotspots within the discipline.

To set the high-frequency keyword collection as meaningful options in discipline

The keywords under consideration fall into the top 1% of all journal papers in a given discipline within the specified time window. These keywords have been sorted by word frequency from largest to smallest. The research subjects encompassed by this collection are indicative of the research focal points of the discipline within a designated time period. In practice, it is not advisable to count high-frequency words with too long a time horizon to avoid statistical errors caused by the transfer of research hotspots. The transfer of research hotspots corresponds to the rhythm of the evolution and development of each discipline; however, the time window should not be too narrow, taking into account the operability. In this study, high-frequency keywords were utilized as variables in lieu of all keywords. The principal rationale for this approach is that high-frequency words are representative, and the changes in their scope and structure can reflect the overall situation of the development of disciplines. The utilization of all keywords may result in the mixing of too much noise data. After testing and comparing, the criterion of 1% of high-frequency words was found to combine both scientific and operability.

To construct the matrix of indicators

The development of a subject area, over time, is facilitated by academic communication, which functions to accumulate and exchange knowledge. Consequently, human cognition of scientific laws and development direction becomes gradually clearer. Assuming the existence of n predetermined possible options for a specific knowledge point within a subject area, it can be posited that in the initial stage, the uncertainty surrounding these options is comparatively pronounced, resulting in a state of heightened confusion regarding knowledge cognition. Conversely, as the process progresses, the uncertainty pertaining to these options undergoes a reduction, thereby facilitating a gradual enhancement in the clarity of knowledge cognition. This progression can be conceptualized as the incorporation of effective information (negentropy) into the knowledge system.

In the framework of information entropy theory, the n preset possible options are regarded as random, and m indicators are used to describe the clarity of each preset option, i.e. to express the probability (Pi) of each option.

The hypotheses proposed in this study suggest that the probability of different predefined options becoming the dominant research direction is subject to change due to the injection of knowledge and information into this disciplinary system. As the future options of this field become gradually clearer and less uncertain, the value of the information entropy state of this disciplinary knowledge system should decrease.

The proposed indicator matrix is thus constructed as follows:

$$F = \begin{pmatrix} f_{11} & \cdots & f_{1i} & \cdots & f_{1m} \\ \vdots & & \vdots & & \vdots \\ f_{i1} & \cdots & f_{ij} & \cdots & f_{im} \\ \vdots & & \vdots & & \vdots \\ f_{n1} & \cdots & f_{nj} & \cdots & f_{nm} \end{pmatrix}$$

Where *n* represents n disciplinary options and *m* represents m indicators. Let (i=1,2,...n; j=1,2,...m), then f_{ij} is the value of the j indicator in the i disciplinary option.

To Select indicators for calculating information entropy of knowledge systems

The selection of indicators is typically undertaken using various methods, including those based on rough set theory, expert research and comment in the field, or the application of correlation coefficient and coefficient of variation methods, among others. Despite the absence of a universally accepted method for indicator screening, the role of expert review in this process remains indispensable.

In this study, the method of expert deliberation is employed for the selection of indicators. The selection process was informed by the study's objective of demonstrating the research volume, extensiveness, activity, and growth capacity of different development directions. It also took into account the scientific and accessible nature of the indicators. Following extensive adjustments and experimentation, and taking into account the research and consultation opinions of peer experts, it was determined that the following seven indicators should be used as journal evaluation guidelines and to calculate the information entropy of the knowledge system.

Table 1. Correspondence table between journal evaluation subsystems and indicators for calculating information entropy of knowledge systems.

Uncertainty level	*Subsystems*	*indicators for calculating information entropy of knowledge systems*
(1) Distribution uncertainty (static level)	A1. Scalability	1) Number of published papers
	A2. Wideness	2) Wide distribution of literature
	A3. Sustainability	3) Number of Funded Papers
(2) Relation uncertainty (dynamic level)	B1. Openness	4) Ratio of international co-authored papers
	B2. Collaboration	5) Number of affiliations per paper
	B3. Competitiveness	6) Growth rate of paper share
	B4. Influence	7) Number of citations per paper

To standardize indicators and calculate their probability

Due to the significant differences in the magnitudes, extreme values, etc. of the different indicators, it is necessary to standardize the transformation of the indicator matrices to form a standardized matrix A:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

where (i=1,2,…n; j=1,2,…m), so that $a_{ij} \in [0,1]$.
The standardized formula is:

$$a_{ij} = \frac{f_{ij} - \min\{f_{1j}, f_{ij}, \ldots f_{nj}\}}{\max\{f_{1j}, \ldots f_{ij}, \ldots f_{nj}\} - \min\{f_{1j}, \ldots f_{ij}, \ldots f_{nj}\}} \quad (1)$$

For indicator j each option probability P_{ij} is defined as:

$$p_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}} \quad (2)$$

where (i=1,2,…n; j=1,2,…m).
While f_{ij} is the minimum value, meaning that a_{ij} is equal to *0* and p_{ij} would have been equal to *0*, assign p_{ij} the value *0.0001* to avoid the problem of *ln(0)* in subsequent calculations and the minimal effect on the overall distribution is negligible.

To calculate the value of information entropy of a knowledge system

The information entropy of this study for each single indicator of the isolated system is calculated by the formula: (Shannon 1963)

$$H_j = K \sum_{i=1}^{n} (p_{ij} \ln p_{ij}) \quad (3)$$

where *i* represents the i of the n options that presuppose unspecified knowledge. (i=1, 2...n);
where *j* represents the j of the m indicators used to characterize uncertainty, which can be viewed as the j subsystems of the knowledge system. (j=1, 2...m);
where *K* is the normalization constant to achieve the calculation results. Since the value range of $\sum_{i=1}^{n}(p_{ij} \ln p_{ij})$ is $[\ln\frac{1}{n}, 0]$, *K* takes the value of $\ln\frac{1}{n}$. Therefore, for a single system, the value of *Hj* is distributed in the range of [0,1]. The case of H_j=0 represents the system is absolutely ordered (only one option, the realization probability is 100%, the realization probability of other options is 0); the case of H_j=1 represents the system is absolutely disordered (the realization probability of all the options is exactly the same);
By calculation, the information entropy state value of each of the m predefined possible options can be derived, then the information entropy of the whole knowledge system is the sum of the information entropy of the m subsystems.

$$H = \sum_{i=1}^{n} H_i \quad (4)$$

Since the numerical distribution of H_i is in the range of [0,1], the numerical distribution of H is in the range of [0,n]. The case of H=0 represents the system absolutely ordered (all 7 subsystems have only one option and the probability of the realization of the same option is 100%, the others are 0); the case of H=n (in this study, n=7 because of 7 subsystems) represents the system absolutely disordered (the probability of all options of all subsystems is exactly the same).

This indicator of H can be regarded as a reflection of the quantity of information and uncertainty inherent within an isolated system. To illustrate this, consider a field of research where there exist two or more divergent perspectives on human understanding of the objective world, or the future trajectory of a specific discipline, of which only a limited number of options can be predicted. At this nascent stage, the probability of the realization of each option is relatively equal, and the uncertainty is pronounced. However, as scientific research progresses, the number of feasible options decreases, thereby reducing uncertainty. Consequently, it can be posited that the probability of realizing a proportion of the possible options increases, while the probability of realizing another proportion of the possible options decreases, thus leading to a decline in uncertainty. This decline in uncertainty can be interpreted as a gradual discernment of the unknown, facilitated by the dissemination of scientific research findings, which in turn leads to a more profound understanding of the objective world by human beings.

To Calculate the discrepancy in information entropy values between factual and counterfactual knowledge systems (negentropy contributed by journal evaluated)

As a background (truth value) for the evaluation of counterfactuals, it is first necessary to calculate the information entropy value H_p of the factual knowledge system for discipline p. For the evaluated journal j as a node in the citation network with in discipline p (Chen 2004), calculate the information entropy value $H_p(j)'$ of the counterfactual knowledge system in the absence of the journal j in discipline p. The change in the values of information entropy in discipline p ($\Delta H_p(j)$) before and after removing of the journal x is the negentropy that the journal j contributes to the knowledge system of discipline p.

$$\Delta \mathrm{H}_p(\mathrm{j}) = \mathrm{H}_p{}'(\mathrm{j}) - \mathrm{H}_p \qquad (5)$$

Result

Information entropy of factual knowledge systems for 112 discipline and their changes along the time dimension

In this study, the list of high-frequency keywords screened based on the papers included in CSTPCD 2016 will be utilized to calculate the annual information entropy values of each discipline in the subsequent database of 2016-2019.CSTPCD 2016 comprised approximately 565 thousand papers, utilizing around 1.5 million keywords and more than 4.1 million times. On average, each paper employed 7.3 keywords. The high-frequency keywords listed in the top 1% in terms of frequency of use for each discipline were calculated. For instance, within the discipline of

"Infectious diseases and infectious diseases", CSTPCD 2016 encompassed eight journals and published 1,093 papers in 2016, utilizing 4,349 keywords and being cited 7,876 times. The 4,349 keywords were then sorted according to their frequency of occurrence, and the 44 keywords that ranked in the top one percent (1% of 4,349) were identified as the set of high-frequency keywords for the discipline.

The information entropy values for each discipline in the database from 2016 to 2019, along with their temporal trends, are shown in Appendix 1.

The analysis of the changes in the information entropy of the knowledge system of each discipline from 2016 to 2019 (see Appendix 1) reveals a clear trend of decrease in entropy values for the majority of disciplines. A comparison of the magnitude of change in the values between 2016 and 2019, as illustrated in Figure 1, reveals that among the 112 disciplinary categories, a mere 11 categories demonstrate an increase in the direction of change in information entropy over the four-year period. The remaining categories, accounting for over 90% of the total, exhibit a decline in information entropy. Given that the numerical comparison of the information entropy of the knowledge system between individual disciplines appears to lack clear significance, the data presented in this study do not provide compelling evidence to support the hypothesis that the information entropy of the knowledge system conforms to a random distribution. This is despite the fact that the distribution state depicted in the figure bears a resemblance to a normal distribution.

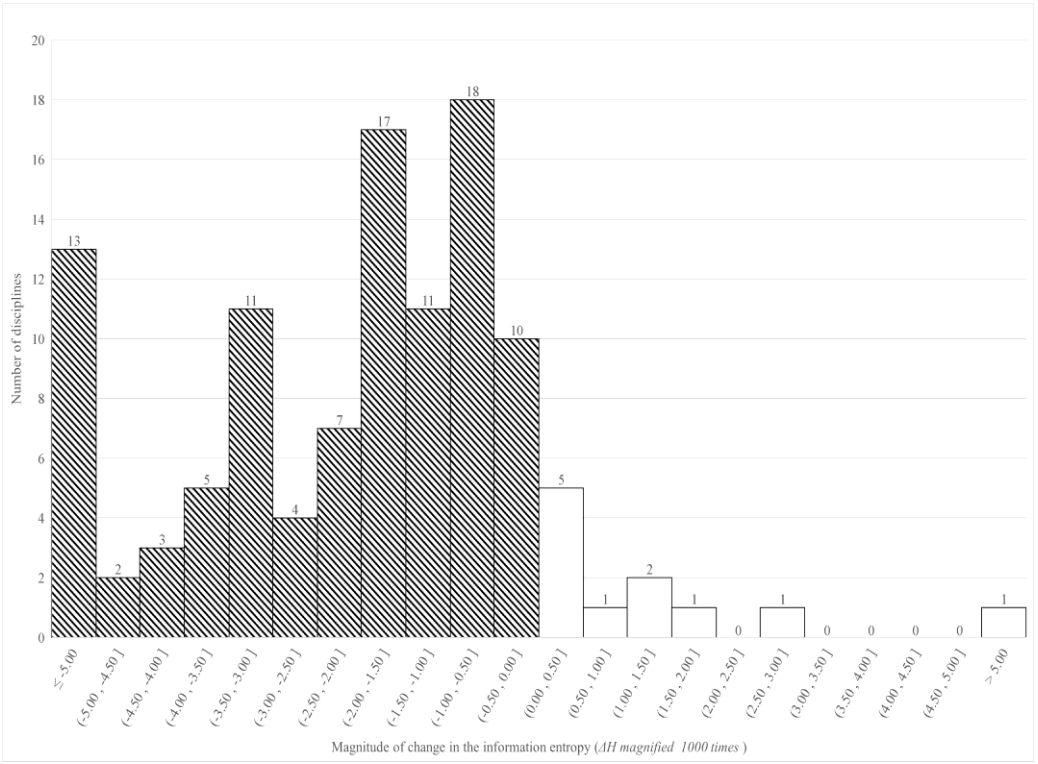


Figure 2. Distribution of the magnitude of change in the information entropy of knowledge systems 2016-2019 for 112 disciplines.

The distribution of changes in information entropy of disciplinary knowledge systems validates the hypothesis of this study: the direction of knowledge system development is evolving from chaos to order with the roles of academic journals which input valuable information and reduce uncertainty in scientific understanding. This shows that the intellectual uncertainty of most disciplines is gradually decreasing, meaning that the thematic direction of the development of a fixed range of disciplines is gradually becoming more focused and clearer, in line with the perception of the general law of disciplinary development. In this process of change, the role played by individual journals varies, that is, the size of the contribution of individual journals varies.

Contribution of evaluated journals to the knowledge system

The contribution of the evaluated journals to the knowledge system can be reflected by calculating the discrepancy in the information entropy (ΔH) between factual and counterfactuals knowledge system. Statistically, the vast majority of the sample has a positive ΔH, with 3,578 journals (96.7%) out of 3,713 journals having a positive ΔH. This indicates that the vast majority of academic journals contribute to the reduction of the chaos of the knowledge system to which they belong, that is, the academic publishing activities of journals fulfill their necessary functions.

According to the information entropy theory, the amount of information introduced into an isolated system should be non-negative, i.e., the most extreme phenomenon is that the amount of information contributed by journals to the system is zero, and the contribution of journals to the system should not have a negative value. However, in this study, the contribution ΔH of some journals to the should-knowledge system to which they belong is negative, which may indicate that these journals have published articles that have a negative effect on the development of the discipline and on the cohesion of the consensus, which increases the degree of confusion in the system.

Since the knowledge system constituted by the collection of papers obtained by using each discipline's high-frequency words as search terms is a mutually independent system in this study, there is no direct comparability between the H state values of different systems, nor between the changes in state values ΔH. However, the direction of ΔH reflects whether journals have positively or a negatively contributed to the system.

In the case of Astronomy, the calculation of the contribution of the six evaluated journals in this discipline to the knowledge system is shown in Table 2.

Table 2. Contributions to the knowledge system (negentropy provided) by six journals in the discipline of astronomy(p) in 2016.

Journal(j)	*Information entropy of factual knowledge system* H_p	*Information entropy of counterfactuals knowledge system* $H_p(j)'$	*Discrepancy as journal's contributions to the knowledge system (negentropy)* $\Delta H_p(j)$
Title 1	15.41	16.09	0.68
Title 2		15.97	0.56
Title 3		15.36	-0.05
Title 4		15.48	0.07
Title 5		15.65	0.24
Title 6		15.50	0.09

Note: The *H* data has been magnified 100 times for ease of display.

In the vast majority of academic disciplines, a small number of journals are found to make a disproportionately large contribution to the overall system in terms of the entropy of information (ΔH) compared to other journals, such as Title 1 and Title 2 in Table 2. The majority of journals, however, have entropy of information (ΔH) values that are almost negligible. This indicates that within the discipline, the distribution of the numerical values of the contribution of many journals to the information entropy of the knowledge system exhibits a distribution pattern with a small number of journals contributing more and a clear long tail of the distribution curve. This finding suggests that only a limited number of journals within the discipline are capable of fulfilling the primary function of academic publications, which is to reduce uncertainty in scientific understanding. Conversely, a greater number of journals have a negligible impact on the reduction of uncertainty in the discipline.

This pattern aligns with Bradford's Law, which posits that a limited number of pivotal core area journals predominate within each discipline. The study revealed that the number of journals contributing substantially to the discipline's knowledge system is also modest, and these journals are designated as "nucleus journals" in a broader sense. However, the study's current limitations preclude the quantification of the relationship between the number of high-contributing journals and the number of low-contributing journals.

The majority of the journals in the sample demonstrate positive ΔH, yet 41 (1.1%) journals exhibit negative ΔH, and 94 (2.5%) journals display 0. When ΔH is 0 or near to 0, it can be deduced that these journals contribute a negligible amount to the development of the discipline. The calculation method employed in this study is predicated on keyword statistics; consequently, journals that are not aligned with the subject matter of the discipline may not be adequately captured, resulting in a contribution value of 0. Additionally, the clarity of the journals' disciplinary classification may be inadequate when the ΔH is negative, resulting in a positive ΔH

for journal classification into discipline p1 and a negative ΔH for journal classification into discipline p2. Another possibility is that the journal publishes content that is too broadly distributed across multiple disciplines. In such cases, the journal's contribution to a specific discipline may be negligible. Statistically, the ΔH of journals is lower in disciplinary categories where synthesis is more pronounced.

Conclusion

In classical information theory, the measurement of the amount of information does not take into account the content importance or intrinsic significance of the information. There is no necessary connection between the amount of information and the importance of the message, and the classical information entropy only calculates a numerical value at the quantitative level, which does not directly indicate the importance of the message. Therefore, in this study, the physical meaning of the indicator values needs further discussion. Particularly for the relatively large number of medium-level journals, the values are less discriminating, leading to deficiencies in areas such as interpretability and assessment of the effectiveness of practice.

The present study operates under the assumption that the journals under review are not currently incorporated within the system. The notion of observing alterations within the system can be conceptualized as a counterfactual analytical approach, which encompasses the formulation of counterfactual assumptions, the establishment of conditions that are antithetical to established facts, and the subsequent measurement of values that are challenging to quantify using conventional descriptive methods. The notion of "counterfactual" research involves the formulation of counterfactual assumptions, the establishment of conditions that are antithetical to the established facts, and the subsequent evaluation of the causal relationship between the change of counterfactual conditions and the results derived from counterfactual reasoning. In the context of complex evaluations of relevant factors, traditional causal analysis frequently assumes that the researcher has controlled the important factors explaining the dependent variable and has not omitted important independent variables. However, the situation and variables under study often fail to satisfy this assumption, or the observed objects are not randomly occurring. This frequently generates endogeneity or sample selection bias, resulting in inaccuracy and bias, or even error, in causal analysis. The advantage of counterfactual analysis is that it can clearly identify differences in baseline or heterogeneity of causal effects among different sample groups that cannot be adequately captured by traditional regression analysis, and then conduct accurate causal analysis.

The methodology employed in this study to define disciplinary knowledge systems utilizes journal categories for classification, a process that may encounter limitations with regard to cross-disciplinary applicability. Future considerations will include the delineation of the boundaries and scope of knowledge systems at the level of the subject matter of the paper, with a view to enhancing the precision and breadth of the application of the methodology.

In this study, there may be limitations in the adequacy of the quantitative results to characterize the reality due to the relatively small number of indicators selected to describe the uncertainty of the system.

For the purpose of data acquisition, the present study employs Chinese literature databases to evaluate Chinese scientific and technical journals. In future, the intention is to adopt international literature databases with more extensive coverage to evaluate international scientific and technical journals.

References

Baldwin, M. (2017). In referees we trust? *Physics Today*, 70, 44-49.

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 2014, 8, 895-903.

Bradford, S.C. (1934). Sources of information on specific subjects. *Engineering an Illustrated Weekly Journal*, 137, 85-86. Reprinted in *Journal of Information Science*, (1995), 10, 176-180.

Chen, C., Hicks, D. (2004). Tracing knowledge diffusion. *Scientometrics*, 2004, 59, 199-211.

Cronin B. (2014). Beethoven vs. Bieber: on the meaningfulness of (alt)metrics. *Libraries in the Digital Age(LIDA) Proceedings*, 13, 15-21. North America.

Dorta-Gonzalez, P., Dorta-Gonzalez, M. I. (2013). Impact maturity times and citation time windows：The 2-year maximum journal impact factor. *Journal of Informetrics*, 7, 593-602.

Editorial Committee of CLC, National Library of China. (2010). *Chinese Library Classification*, Beijing: National Library Press.

Eysenbach, G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 2011, 13, e123.

Feng, S., Li, H., Qi, Y. (2023). How to detect the sleeping beauty papers and princes in technology considering indirect citations? *Journal of Information Science*, 17.

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 1955, 122, 108-111.

Huang, Y., Li R., Zhang, L., Sivertsen G. (2021) A comprehensive analysis of the journal evaluation system in China. *Quantitative Science Studies*, 2, 300–326

Kahneman, D., Slovic, P., Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*, Cambridge: Cambridge University Press.

Karanatsiou, D., Misirlis, N., Vlachopoulou, M. (2017). Bibliometrics and altmetrics literature review Performance indicators and comparison analysis. *Performance Measurement and Metrics*, 18, 16-27

Lewis, D.W. (1989). Economics of The Scholarly Journal. *College & Research Libraries*, 50, 674-688.

Ma, L. (2008). Comparative analyses on full-text databases of Chinese journals. *Journal of Nanjing Institute of Industry Technology*.

Ma Z. (2022). A Counterfactual Method for Evaluating Academic Journals through Changes in Information Entropy in Knowledge Systems. *Journal of the China Society for Scientific and Technical Information*, 41,741-765.

Ma, Z., Pan, Y., Wu, Y., Yu, Z., & Su, C. (2022). Measuring the Competitive Pressure of Academic Journals and the Competitive Intensity within Subjects. *ArXiv*, https://arxiv.org/abs/2211.17164 .

Maflahi, N, Thelwall, M. (2016). When are readership counts as useful as citation counts? Scopus versus Mendeley for LIS journals. *Journal of the Association for Information Science and Technology*, 67, 191-199.

Narin, F., Hamilton, K.S. (1996). Bibliometric performance measures. *Scientometrics*, 36, 293-310.

Pan, Y., Ma, Z. (2018) *Chinese Science and Technology Journal Citation Reports (Core Edition)*, Beijing: Scientific and Technical Documentation Press.

Preim, J., Taraborelli, D., Groth, P., et al. (2010). *Altmetrics: A Manifesto*. Retrieved January 9, 2025 from: https://zenodo.org/records/12684249 .

Priem, J., Piwowar, H. A., Hemminger B. M. (2012). Altmetrics in the wild: using social media to explore scholarly impact. *ArXiv*, https://arxiv.org/abs/1203.4745 .

Shannon, C.E, Weaver, W. (1963) *The mathematical theory of communication*. Champaign: University of Illinois Press.

Shuai, X., Pepe, A., Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, Twitter mentions, and citations. *PLoS ONE*, 2012, 7, e47523.

State Bureau of Quality and Technical Supervision of PR China. (2009) *National Standard of PRC: Classification and Code of Disciplines (GB/T 13745-2009)*, Beijing: China Quality and Standard Publishing.

Thelwall, M., Haustein, S., Lariviere, V., et al. (2013) Do altmetrics work? Twitter and ten other social web services. *PLoS ONE*, 2013, 8, e64841.

Vinkler, P. (2009). Introducing the Current Contribution Index for characterizing the recent, relevant impact of journals. *Scientometrics*, 79, 209-420.

Wang, Z., Wang, X., Ma, J. (2011). Scientificalness and limitations of Index as a tool for bibliometric analysis. *Journal of the China Society of Indexers*, 33-41.

Zhou, P., Leydesdorff, L. (2007). A comparison between the China Scientific and Technical Papers and Citations Database and the Science Citation Index in terms of journal hierarchies and interjournal citation relations. *Journal of the American society for Information Science and Technology*. 58, 223-236.

Appendix

Correspondence table between journal evaluation subsystems and indicators for calculating information entropy of knowledge systems.

Discipline(p)	*Number of options(n) (high-frequency keywords) in 2016*	*Factual value of information entropy H_p in 2016*	*Factual value of information entropy H_p in 2017*	*Factual value of information entropy H_p in 2018*	*Factual value of information entropy H_p in 2019*
Multidiscipline	182	86.55	82.07	81.32	83.69
General University Journal	532	148.62	148.24	142.93	147.21
Normal University Journal	230	60.16	61.77	63.35	61.74
Mathematics	200	30.68	31.36	30.31	28.96
Information Science and System Science	130	33.39	31.48	30.16	29.63
Mechanics	188	9.27	9.19	8.95	8.89
Physics	392	34.23	32.13	30.86	29.85
Chemistry	406	42.49	42.96	41.07	40.74
Astronomy	29	15.41	14.40	13.81	13.16
Earth Science	110	27.71	28.29	27.30	27.52
Atmospheric Sciences	101	16.00	15.58	14.77	14.23
Geophysics	162	36.93	36.51	34.26	35.12
Geography	210	24.27	23.85	23.76	22.89
Geology	261	77.90	73.06	72.08	72.43
Marine Science, Hydrography	173	51.02	52.18	50.92	47.71
Basic Biology	199	54.82	53.94	51.14	49.76
Ecology	164	29.44	29.04	28.89	27.81
Botany	92	17.78	17.99	17.79	16.89
Entomology, Zoology	83	28.11	27.16	26.11	25.06
Microbiology, Virology	92	24.72	24.61	24.84	23.82
Psychology	76	19.75	19.36	19.83	18.78
Agribusiness	557	156.32	154.65	150.90	146.34
Agricultural University Journal	305	79.41	78.79	80.56	76.20
Agronomy	152	46.95	45.13	44.40	44.95
Horticulture	92	10.49	9.96	10.16	9.80
Soil Science	70	22.69	22.67	23.05	23.15
Plant Protection	82	18.68	19.14	18.37	17.56
Forestry	225	20.42	21.01	19.76	20.27
Animal Husbandry, Veterinary Science	200	21.61	20.92	21.08	19.69

Discipline(p)	*Number of options(n) (high-frequency keywords) in 2016*	*Factual value of information entropy H_p in 2016*	*Factual value of information entropy H_p in 2017*	*Factual value of information entropy H_p in 2018*	*Factual value of information entropy H_p in 2019*
Grassland Science	60	13.67	13.02	12.78	12.64
Aquaculture	182	31.67	32.43	32.07	32.08
General Medicine	668	111.49	109.92	111.71	108.25
Medicine and Pharmacy University Journal	541	164.93	166.04	155.62	159.79
Basic Medicine	230	57.48	54.26	51.45	51.65
Clinical Medicine	452	108.06	108.10	104.23	97.47
Clinical Diagnostics	162	40.05	41.12	41.65	40.24
Health Care Medicine	106	33.14	31.93	29.92	29.98
Internal Medicine	37	18.43	17.42	16.59	17.06
Cardiovascular Disease	152	20.17	19.44	19.11	17.79
Respiratory Disease, Tuberculosis	63	18.87	17.71	18.21	17.44
Gastroenterology	100	30.50	30.84	29.96	29.31
Hematologic, Nephrology	86	20.35	20.95	19.94	18.69
Endocrinology and Metabolic Disease, Rheumatology	53	13.46	13.29	13.56	13.21
Infectious Diseases, Infectious Diseases	44	15.00	14.92	14.96	14.18
Comprehensive Surgery	148	54.97	51.18	51.90	49.61
General Surgery, Thoracic Surgery, Cardiovascular Surgery	134	44.43	43.79	41.00	39.23
Urology	47	11.05	10.79	10.77	10.19
Orthopaedic Surgery	82	19.97	20.11	19.51	19.03
Burn Surgery, Plastic Surgery	72	25.32	26.00	26.72	25.03
Obstetrics and Gynaecology	69	15.68	15.65	15.09	14.34
Paediatrics	118	18.44	17.62	17.51	17.66
Ophthalmology	95	16.49	15.73	16.16	15.99
Otolaryngology	79	22.72	21.15	21.67	21.79
Stomatology	125	36.97	37.82	38.30	36.21
Dermatology	64	13.42	13.67	13.37	12.80

Discipline(p)	*Number of options(n) (high-frequency keywords) in 2016*	*Factual value of information entropy H_p in 2016*	*Factual value of information entropy H_p in 2017*	*Factual value of information entropy H_p in 2018*	*Factual value of information entropy H_p in 2019*
Sexual Medicine	51	9.06	8.73	8.35	7.83
Neurology, Psychiatry	171	28.64	29.37	29.25	29.98
Nuclear Medicine, Medical Imaging	176	34.53	32.93	31.14	29.86
Oncology	200	44.64	42.31	41.66	41.98
Nursing	205	34.03	34.89	32.46	33.01
Preventive Medicine and Public Health	194	28.83	28.19	27.17	25.71
Epidemiology, Environmental Medicine	235	68.93	64.31	60.29	57.08
Eugenics	114	18.36	18.38	17.29	16.47
Health Management, Health Education	57	32.73	31.59	29.91	30.27
Military Medicine and Specialty Medicine	306	7.81	7.38	7.34	6.98
Pharmacy	537	116.44	114.70	117.53	119.07
Traditional Medicine	490	57.71	59.32	60.13	58.69
Tradition Medicine University Journal	184	30.35	30.55	29.53	29.75
Integrative Medicine	159	26.70	26.54	26.71	26.97
Traditional Chinese Medicine	527	80.75	79.77	74.25	73.40
Acupuncture and Moxibustion, Orthopaedics and Traumatology	57	5.39	5.38	5.39	5.28
Basic Science for Engineering and Technology	353	21.70	20.94	21.39	21.30
Engineering and Technology University Journal	959	229.03	227.63	224.11	228.55
Information and System Science Related Engineering and Technology	361	67.62	63.93	64.64	62.58
Bioengineering	84	23.52	22.77	21.85	20.55

Discipline(p)	*Number of options(n) (high-frequency keywords) in 2016*	*Factual value of information entropy H_p in 2016*	*Factual value of information entropy H_p in 2017*	*Factual value of information entropy H_p in 2018*	*Factual value of information entropy H_p in 2019*
Agricultural Engineering	324	74.10	74.29	76.17	74.58
Biomedical Engineering	132	16.43	15.42	15.01	14.37
Surveying and Mapping	175	20.86	19.64	18.45	17.39
Materials	282	73.76	73.90	71.69	72.85
Metallic Materials	213	19.75	19.13	19.01	19.52
Mining Engineering Technology	290	67.14	64.34	65.51	66.26
Metallurgical Engineering Technology	113	31.13	30.07	28.33	26.81
Mechanical Engineering Design	444	38.21	37.92	36.40	36.39
Mechanical Manufacturing Process and Equipment	380	48.60	47.41	45.37	46.35
Power Engineering	131	17.23	16.57	15.73	15.93
Electrical Engineering	514	112.97	108.87	103.52	104.77
Energy	249	44.09	42.07	39.81	39.54
Oil and Gas	330	104.23	107.04	109.40	111.46
Nuclear	75	24.08	22.70	21.80	21.07
Electronic	518	33.54	32.05	32.17	31.82
Optoelectronics and Laser	221	30.43	30.33	29.81	28.35
Communication	171	37.08	36.15	36.37	35.53
Computer	706	32.60	33.51	32.99	33.72
Chemical Engineering	399	66.99	67.54	67.23	63.10
Polymer	107	35.57	34.98	34.41	32.15
Fine Chemical Engineering	119	19.67	19.00	17.74	18.04
Applied Chemical Engineering	94	18.14	17.98	18.39	17.62
Instrumentation	249	29.41	27.86	27.72	26.32
Defence	223	50.74	50.20	49.36	46.45
Textile	95	24.88	24.26	24.22	23.10
Food	401	63.39	63.76	63.32	59.71
Building	373	33.56	32.40	32.62	31.93
Civil Engineering	129	16.43	15.74	15.25	14.74
Water Resources Engineering	285	77.99	77.05	75.41	71.06

Discipline(p)	*Number of options(n) (high-frequency keywords) in 2016*	*Factual value of information entropy H_p in 2016*	*Factual value of information entropy H_p in 2017*	*Factual value of information entropy H_p in 2018*	*Factual value of information entropy H_p in 2019*
Transportation Engineering	90	13.34	12.67	12.68	12.52
Road Transportation	135	16.08	15.19	15.06	15.47
Railroad Transportation	129	20.81	19.75	18.59	18.97
Waterway Transportation	157	26.85	26.87	26.75	25.09
Aviation, Aerospace	357	91.56	92.32	91.97	88.19
Environmental and Resource	434	61.06	62.67	59.62	57.45
Safety	150	25.20	24.50	24.02	22.52
Management	298	67.45	64.83	62.42	63.77

Note: The *H* data has been magnified 100 times for ease of display.

https://doi.org/10.51408/issi2025_084

Examining the Cognitive Gap Between Authors and Peer Reviewers on Academic Paper Novelty

Chenggang Yang[1], Chengzhi Zhang[2]

[1]*ichigo@njust.edu.cn*, [2]*zhangzc@njust.edu.cn*

Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094 (China)

Abstract

Novelty is a crucial metric for assessing the quality of academic papers. Scholars strive to highlight the novel aspects of their work, particularly in the title, abstract, and introduction, where they often emphasize the novel contributions of their research. Peer review, serving as the gatekeeper of scientific rigor, rigorously evaluates the innovativeness of papers to ensure they meet the standards of scientific publication. However, there may be a cognitive gap between the self-promotion by authors and the evaluation of novelty by peer reviewers. To investigate whether such a gap exists, we analyzed 15,328 academic papers published in Nature Communications from 2016 to 2021, along with their peer review comments. We extracted promotional statements from the introduction of these papers and evaluative statements on novelty from the review comments, categorizing them into theoretical innovation, methodological innovation, and result innovation. The findings reveal that both reviewers and authors place greater emphasis on result innovation, with reviewers adopting a more comprehensive approach when evaluating novelty. By examining the impact of promotional intensity on reviewers' evaluations in relation to the paper's inherent novelty, we found that highly innovative papers benefit from using more promotional language, receiving more positive evaluations from reviewers. In contrast, excessive promotional language in less innovative papers leads to lower evaluations of their novelty. Based on these results, we suggest that highly innovative papers can enhance positive reviewer evaluations by moderately employing promotional language, while less innovative papers should exercise caution to avoid being perceived as overstating their contributions. Additionally, the study underscores the need for clearer review standards to help reviewers evaluate the innovativeness of papers more objectively, minimizing the influence of promotional language.

Introduction

In recent years, the number of academic papers has grown exponentially. To ensure that high-quality research is published promptly and accurately in appropriate journals or conferences, the pressure on peer reviewers, who serve as gatekeepers of scientific publishing, has intensified. However, peer review is not without its

challenges, including inefficiencies and potential biases(Parker et al., 2018; Stelmakh et al., 2019; Wicherts, 2016). Novelty, as a critical component of academic paper quality, is a key criterion reviewers use to make recommendations for acceptance or rejection.

The ability to communicate novel ideas and research findings effectively is an indispensable part of academic research and is crucial across many scientific domains, such as grant applications, patent writing, and academic paper writing (Peng et al., 2024). Academic papers are the primary medium for disseminating research outcomes, enabling researchers to share their discoveries and insights. To accurately and efficiently convey the innovative aspects of their work, researchers often highlight their contributions in the title, abstract, and introduction of their papers. These promotional statements have been shown to correlate with the subsequent impact of the papers (Pearson, 2020; Wheeler et al., 2021).

Table 1. Examples of three promotion types.

Promotion Type	Author description	Reviewer Comments
Exaggeration	This groundbreaking study introduces a revolutionary method that will completely transform data privacy in machine learning.	The claim that this method will 'revolutionize data privacy in machine learning' appears overly ambitious. For instance, Smith et al. (2021) demonstrated that while advancements in data privacy are significant, they often come with trade-offs in model performance and complexity. I recommend that the authors provide a more balanced view that acknowledges these trade-offs and the context in which their method may be effective.
Insufficient promotion	This study presents a approach to enhance data privacy protection in machine learning. While the method has demonstrated some effectiveness on certain datasets, it has not yet undergone extensive empirical validation.	To avoid understatement, more information about the advantages of the research method, specific experimental results, and its potential impact should be included in the description.
Appropriate promotion	This study presents a novel approach that demonstrates notable improvements in data privacy within machine learning frameworks compared to existing methods. Our results indicate enhanced protection of sensitive information while maintaining model performance.	The authors successfully provide a balanced description of their contributions, clearly articulating the improvements over existing methods without relying on hyperbolic claims.

Note: The parts marked in red are the promotional language used by the author.

However, inappropriate promotion in academic papers can lead to adverse consequences. Some scholars, driven by utilitarian motives or insufficient research of prior studies, may exaggerate the novelty of their work. If the research findings are later proven to be less novel than claimed, not only can the authors' academic reputations suffer significant damage, but other researchers may also be misled, investing time and resources in misguided directions. This can hinder the progress of the entire field and stifle genuine innovation. Conversely, some scholars may insufficient promote or inaccurately describe the novelty of their research, leading to their findings being overlooked and limiting their dissemination within academia and related fields. To prevent such scenarios, peer reviewers, as gatekeepers of scientific quality, rigorously evaluate the merits of submitted papers. Inappropriate promotion can result in setbacks during the peer review process. As shown in Table 1, if exaggeration is detected, reviewers may point out the exaggerations in their comments, such as "*Smith et al. (2021) demonstrated that while advancements in data privacy are significant, they often come with trade-offs in model performance and complexity.*", and provide corresponding references as evidence for the authors to revise their papers. In severe cases, the paper may even be rejected for publication. For papers that insufficient promotion, reviewers may struggle to grasp the innovative aspects of the research, leading to the findings being undervalued. This can result in a lower overall evaluation of the paper's quality by reviewers, ultimately affecting its chances of publication.

Therefore, appropriate promotion is crucial for reviewers to make informed decisions regarding acceptance or rejection. Specifically, we address the following three questions:

Firstly, to investigate in greater detail how authors and reviewers evaluate the novelty of academic papers, we adopted the classification framework proposed by Leahey et al.(2023). This framework categorizes innovation in academic papers into theoretical innovation, methodological innovation, and result innovation. Correspondingly, we classify the evaluations of authors and reviewers into theoretical innovation evaluation, methodological innovation evaluation, and result innovation evaluation. Based on this, we propose RQ1:

RQ1: Which aspects of innovation do paper authors and reviewers prioritize more?

Based on RQ1, we can statistically analyze which aspects of innovation authors and reviewers emphasize more when promoting or evaluating the novelty of academic papers. Building on this, we aim to explore the cognitive differences between authors and reviewers regarding the perceived innovative contributions of papers.

Specifically, we seek to identify which innovative points, after being promoted by authors, are also endorsed by reviewers. This leads us to propose RQ2:

RQ2: What are the differences in focus between paper authors and peer reviewers regarding the innovation of a paper?

Both RQ1 and RQ2 investigate the innovative aspects of papers, but what is the relationship between the intensity of promotion and the evaluation by reviewers? Could it be that the more promotional language authors use, the more positive feedback they receive from reviewers? To study the intensity of promotional language and to prevent both over-promotion and insufficient promotion, we have integrated the novelty indicator Novelty_U proposed by Uzzi et al.(2013), leading us to propose RQ3:

RQ3: What is the relationship between the intensity of promotion and the reviewers' evaluation of novelty?

The primary contributions of this paper are manifested in the following three aspects: Firstly, we have developed a novel methodology for extracting innovation evaluation sentences from academic papers and peer review comments. As an increasing number of journals opt to open their peer review comments, the importance of batch extracting information from a vast amount of peer review data has become more pronounced. Unlike academic papers, peer review comments lack a unified writing standard, making it challenging to extract innovation-related evaluations from them. This study has devised a "rule-based + machine learning" approach that can accurately extract reviewers' evaluations regarding the innovation of papers from peer review comments, thereby contributing to the comprehension of peer review feedback.

Secondly, we have investigated the cognitive biases between paper authors and reviewers regarding the innovative aspects of papers. We began by analyzing which aspects of innovation are prioritized by authors and reviewers during the writing and reviewing processes, respectively. We then observed whether reviewers acknowledged the innovative contributions as described by the authors in each paper, thereby providing a preliminary exploration into the current state of cognitive biases between authors and reviewers.

Lastly, we have examined the relationship between the intensity of promotional language in the introduction of a paper and the level of agreement it receives during the review process. The consequences of using different degrees of promotional language in the introductions of academic papers, and whether more promotional language is invariably better, remain largely unexplored in current research. This study contributes to the understanding of these dynamics, thereby advancing the

cause of reasonable promotion in the writing of academic papers.

Related work

Current research on promotional language in academic papers predominantly focuses on the titles and abstracts, while studies on peer review seldom address the extraction of innovation evaluation sentences. The related work section of this study encompasses two parts: research on promotional language in academic papers and research on innovation evaluation in peer reviews.

Innovative promotional language

Promotional language refers to the linguistic expressions and stylistic choices designed to market or advocate research findings. This type of language is often characterized by exaggeration, subjectivity, or emotional appeal, which may influence the reader's objective understanding of the research. Previous studies on promotional sentences have primarily concentrated on grant or project proposals. For instance, Millar et al.(2022) analyzed 717 NIH grant applications and found that applicants increasingly describe their work subjectively, relying on promotional language and emotional appeals. Peng et al.(2024) examined the promotional language in funding applications from NIH, NSF, and the Nord Foundation, investigating its relationship with the likelihood of funding and the future impact of the projects. It is evident that the dissemination of research not only depends on the output of research results but is also closely related to the manner in which it is promoted.

Current research on promotional language in academic papers delves into the titles, abstracts, and main text content, aiming to uncover phenomena present in scientific research and to provide recommendations for academic writing, thereby enhancing the quality of scholarly articles. Citation counts are often used as a proxy for the influence of a paper to study the effects of promotional language. Titles and abstracts, serving as summaries of the entire paper, are common corpora for research on promotional language in papers. Metrics such as length, vocabulary usage, and semantic complexity have been extensively studied by many researchers(Jiang & Jiang, 2023; Li, 2022; Pearson, 2020; Sagi & Yechiam, 2008; Wheeler et al., 2021), as detailed in Table 2. However, the main text of academic papers remains under-researched due to the challenges in data acquisition and processing. In recent years, the open access to large-scale paper datasets and the development of large language models have propelled research in the semantic understanding of full-text academic papers. Such research often correlates writing style with other metrics. For example,

Lu et al.(2019) found that cultural background influences sentence structure and word choice. Costello et al.(2023) discovered a relationship between gender and writing style, as well as the crucial role of editors in mitigating these differences, by examining the use of uncertain language in papers written by male and female authors. Wu et al.(2025) found that the introduction part of the paper is more suitable for measuring its novelty.

Moreover, there is relatively scant research utilizing large-scale corpora to investigate the use of promotional language in academic papers. Thanks to the advancements in large language models, we are now better equipped to process the vast corpora that are currently available, extracting more information from them. This study draws on research related to promotional language in grant applications, employing large language models to automatically extract and classify promotional language from the introductions of academic papers. Based on this, we assess the intensity of promotional language in academic papers and examine the relationship between the intensity of promotion and the level of endorsement by reviewers.

Table 2. Related works of promotion language in academic article.

Source	Authors	Contribution
Title	Pearson(2020)	The study found that the structure and characteristics of titles may influence a paper's academic impact.
	Sagi & Yechiam(2008)	The study provides empirical evidence that humorous titles in scientific articles are associated with fewer citations.
	Jiang & Jiang, (2023)	Reveal significant trends in title length, complexity, and syntactic structures.
Abstract	Wheeler et al.(2021)	Reveal a significant increase in the use of personal pronouns and expressive confidence (referred to as "clout") in psychology journal abstracts.
	Li(2022)	It explores the relationship between passive voice usage and active voice initiated by personal pronouns, contributing to a better understanding of the evolving style of academic writing.
	Song et al.(2023)	The findings suggest that papers published in higher quartile journals tend to exhibit greater lexical density and sophistication, implying a connection between writing quality and scientific impact.

Innovation Evaluation in peer review

Peer review stands as the cornerstone of academic exchange and the bedrock of scientific publishing(Ghosal et al., 2022). Peer experts, with their profound domain knowledge and professional experience, are capable of evaluating the overall quality of academic papers. The innovativeness of a paper, being a decisive factor of its quality, has always been highly regarded by reviewers(Teplitskiy et al., 2022).

However, the peer review process has been subject to controversy due to its lengthy review cycles, lack of transparency, and potential biases (Parker et al., 2018; Stelmakh et al., 2019; Wicherts, 2016). To address these issues, several journals, including Nature Communications and Plos One, have begun to make peer review comments publicly available ("Transparent Peer Review for All," 2022). The disclosure of peer review comments has promoted transparency in the review process and enhanced the efficiency of communication between paper authors and reviewers. Reviewers are expected to make rational judgments about the quality of papers and provide revision suggestions to authors based on these judgments, without being influenced by other factors. For instance, Sun et al.(2024) analyzed peer review comments from Nature Communications and found that authors who used the second person in their communications with reviewers received more positive evaluations.

To enhance researchers' understanding of innovation evaluation during the review process, we employed rule-based and machine learning methods to extract innovation evaluation sentences from peer review comments. We studied the current methods and focus points of innovation evaluation in the review processes across different disciplines and, in conjunction with the promotional language in the introductions of papers, provided recommendations for authors on the use of promotional sentences in academic writing.

Data and Methodology

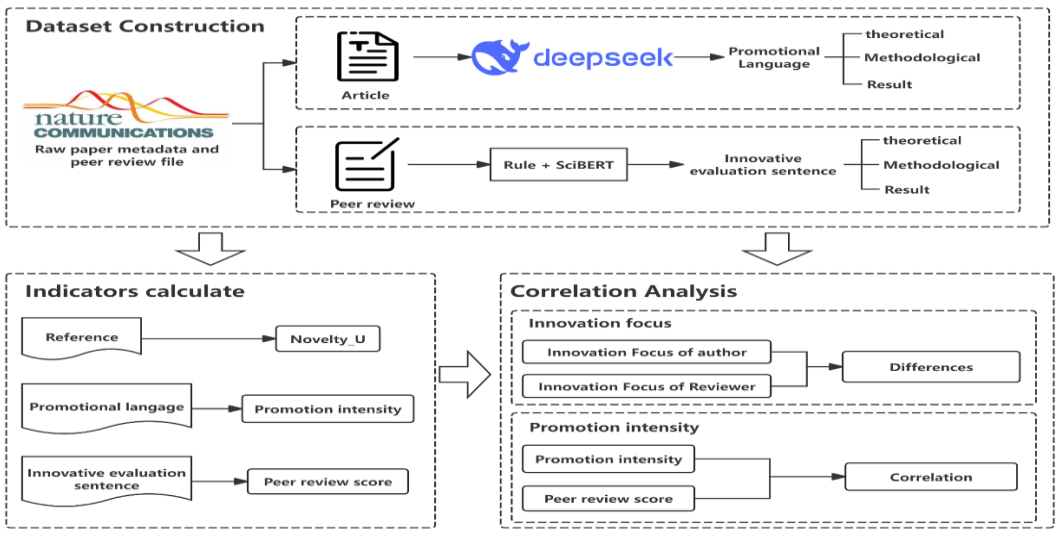


Figure 1. Framework of this study.

The aim of our study is to examine the differences in focus between paper authors and reviewers regarding the novelty of papers during the publication process, as well as the relationship between the use of promotional language by authors and the evaluations by reviewers. To achieve this, we utilized original academic papers and peer review comments from various fields in *Nature Communications*, extracting sentences related to innovation evaluation for analysis. We also assessed the promotional intensity of the original papers using the novelty metric proposed by Uzzi et al.(2013). The framework of our study is illustrated in Figure 1. Specifically, we conducted our research in three steps. The first step involved the construction of the dataset, where we collected all academic papers published in *Nature Communications* from 2016 to 2021 along with their publicly available peer review comments. We parsed their contents to extract authors' promotional language about their own research from the introduction of original papers and reviewers' innovation evaluation sentences from the review comments. The second step was the comparison of innovation focus points across different disciplinary fields. Based on the five disciplinary categories provided by the *Nature Communications* website, we observed how researchers' focus on innovation varies across fields and how the focus points of paper authors and reviewers differ regarding the novelty of papers. The third step combined a reference-based method for calculating paper novelty to investigate the relationship between the use of promotional language in papers with different levels of novelty and reviewer comments.

Dataset and Data Preprocessing

This section outlines the process of dataset construction and preprocessing for our study. Initially, we collected the original papers and peer review comment files from the *Nature Communications*[1] website. Subsequently, we employed large language models to extract promotional language from the original papers and utilized a "rule-based + machine learning" approach to extract innovation evaluation sentences from the peer review comments. This groundwork lays the foundation for our subsequent analysis of the authors' and reviewers' evaluations of the papers' innovativeness.

Raw article and peer review corpus collection: The data source for this study is Nature Communications, a subsidiary journal of Nature. This journal encompasses the latest research findings across various fields of natural sciences and has been committed to the transparency of peer review to enhance the quality of the review process, being one of the earliest journals to make peer review comments publicly available. Since 2016, authors have had the option to disclose the exchanges between

[1] https://www.nature.com/ncomms/

themselves and the reviewers. Papers with disclosed review comments can be found with corresponding peer review PDF files on their content pages, which include the reviewers' comments and the authors' responses from each round of review.

We collected the publication dates, titles, abstracts, main texts, and publicly available peer review PDF files of all papers published from 2016 to 2021 from the *Nature Communications* website. Based on the journal's disciplinary classification, we categorized the papers into five fields: biological science, health science, earth and environmental science, physical science, and scientific community and society. A single paper could belong to multiple disciplinary fields. We identified the structure of the papers using HTML tags within the main text and extracted the introduction sections as the corpus for subsequent promotional language extraction. We required that the collected papers have complete titles, authors, abstracts, and introduction content, along with publicly available peer review comments. Ultimately, we gathered 15,328 academic papers along with their peer review comments. Since the reviewers' comments and authors' responses in the publicly available peer review comments for each paper were contained within the same PDF file, we used the Python package PyMuPDF [2] to parse the text and segmented the reviewers' comments from the authors' responses based on linguistic features and font size characteristics.

Extract promotional language from academic papers: To investigate the promotional intensity of authors regarding the innovative aspects of their work, we need to extract contribution-promoting sentences from the papers. Here, we define contribution-promoting sentences in academic papers as "sentences used to explicitly highlight the main contributions and innovative points of the research work." Although authors tend to emphasize the key points of their research in the title and abstract, they may still lack descriptions of innovative aspects in some detailed parts. In the introduction of a paper, authors clarify the research background while explicitly stating the specific problems to be solved or the core themes of the research, and they articulate the purpose and significance of the study, promoting the main innovative points of the research. Therefore, we selected the introduction of the paper as the corpus for extracting contribution-promoting sentences, extracting these sentences from the title, abstract, and introduction of the paper. Additionally, referencing the classification method of academic paper innovation by Leahey et al.(2023), we categorized the innovation in academic papers into theoretical innovation, methodological innovation, and result innovation, with specific definitions of each type of innovation as shown in Table 3.

[2] https://pymupdf.readthedocs.io/en/latest/

Table 3. Definition of three Innovation types.

Innovation Type	Definition
Theoretical Innovation	Refers to breakthroughs in theoretical frameworks, models, or concepts. This can involve new theoretical perspectives, redefinitions of concepts, or extensions of existing theories, which advance the understanding and development of the discipline.
Methodological Innovation	Involves improvements or innovations in research methods, techniques, or tools. This can include new experimental designs, data collection methods, and analytical techniques, making research more efficient and reliable, or enabling the resolution of previously unsolvable problems.
Result Innovation	Refers to new findings or conclusions obtained from the research. This type of innovation emphasizes the new knowledge or data gained from the research and its potential applications, which can have a significant impact on theory, practice, or policy.

When extracting contribution-promoting sentences from academic papers, we opted to utilize a large language model for this task. We employed DeepSeek-V3[3] as our extraction model. DeepSeek-V3 is an exceptional Mixture of Experts (MoE) language model with an overall parameter scale of 671B, where each token activates 37B parameters, and it has surpassed other open-source models in performance across multiple test datasets (DeepSeek-AI et al., 2024). We referenced the prompt templates provided by DeepSeek's official documentation to craft corresponding prompts that define innovative contribution sentences, instructing the large model to extract the original text of innovative contribution sentences from the titles, abstracts, and introductions of papers. To investigate the innovative points that academic paper authors focus on, we directed the large model to extract contribution-promoting sentences and categorize them into theoretical innovation, methodological innovation, and research outcome innovation, with each contribution-promoting sentence belonging to only one category. After completing the prompt, we tested its extraction performance, randomly selecting 10 papers after each extraction test to observe the results, ensuring that all contribution-promoting sentences in the papers were extracted and assigned to the correct category, and that these sentences were sourced from the original text rather than generated by the model. Once the extraction performance met our expectations, we used the refined prompt to extract

[3] https://www.deepseek.com/

and classify contribution-promoting sentences from the introductions of papers across the entire dataset. The final prompt we used is shown in Table 4.

Extract innovative evaluation sentences from peer review: To investigate reviewers' opinions on academic papers, we developed a "rule-based + machine learning" method to extract innovation evaluation sentences from review comments. Initially, we referenced the work of Leahey et al.(2023) on extracting innovation evaluation sentences from academic papers, using a large language model to generate common templates for innovation evaluation in peer review comments. We also conducted a survey of language patterns in existing peer review corpora to develop and refine a lexicon for the preliminary extraction of innovation evaluation sentences, as detailed in Table A1 in the Appendix. However, due to the polysemous nature of innovation indicator words in peer review comments—for example, 'original' can mean innovative or refer to the reviewer's initial opinion when placed before 'review'—we established corresponding rules. For instance, if 'new' is followed by words such as 'version', 'fig', or 'review', it is not recognized as a candidate for an innovation sentence. Based on this lexicon and these rules, we used regular expressions to extract candidate innovation sentences from peer review comments, initially extracting 108,033 sentences containing innovation indicators from 15,328 peer review comments.

To address the issue of low recall in the rule-based extraction method, we employed a machine learning approach to further classify the initially extracted innovation evaluation sentences. Specifically, we utilized SciBERT as our classification model to determine whether the preliminarily extracted innovation evaluation sentences were genuinely related to innovation. SciBERT is a pre-trained model based on the BERT architecture, optimized specifically for scientific texts and currently applicable to various natural language processing tasks such as text classification, named entity recognition, and question-answering systems, particularly in scientific applications (Beltagy et al., 2019). We randomly selected 1,500 candidate innovation evaluation sentences for manual annotation, distinguishing between innovation evaluation sentences and non-innovation evaluation sentences. The annotation was performed by two graduate students in library and information science, with a unified definition of innovation evaluation sentences in peer review comments as "sentences in which reviewers make positive or negative evaluations about the innovation of the paper's theory, methods, results, etc." The Kappa coefficient calculated after annotation was 0.92. We then divided all 1,500 sentences into training, validation, and test sets in an 8:1:1 ratio to evaluate the model's performance on this task. Ultimately, our extraction model achieved an accuracy of 0.88, a recall of 0.94, and

an F_1 score of 0.92 on the test set.

We employed SciBERT to classify the pre-extracted candidate sentences into two categories: innovation evaluation sentences and non-innovation evaluation sentences. Ultimately, we extracted 38,561 innovation evaluation sentences from all peer review comments and categorized them into three types: theoretical innovation, methodological innovation, and result innovation. The classification rules are detailed in Tables A2, A3, and A4 in the Appendix, and the extraction results are illustrated in Figure 2. Not every innovation evaluation sentence falls into one of these three categories, as some sentences provide an overall assessment of the paper's innovativeness. For example, the sentence "*In my opinion, the novelty of this work is enough to guarantee a publication in Nature Communications.*" expresses a positive evaluation of the paper's overall innovativeness without specifying any particular aspect of innovation.

Table 4. Prompt used in promotion language extraction.

Instruction	**The persona pattern:** You are a proficient linguist skilled in reading academic articles. **Introduce the target of our task:** You will be given a paragraph in the Introduction section of a publication. Please follow the instructions to label the paragraph in the Introduction section provided by user. In the introduction of a paper, the author mentions the innovations of the entire work, which we define as contribution statements. **Definition of contribution statement:** Contribution statements are sentences or paragraphs in academic papers that clearly highlight the main contributions and innovations of the research work. **Definition of three types of innovation:** These contributions can be categorized into three types: **(See Table 3.)** **Introduce the details in our task:** Each paper's introduction may contain one or more of these three types of innovations, and each sentence belongs to only one type of innovation. Please read the provided research paper's introduction carefully and extract the original sentences representing these contribution statements, categorizing them into theoretical innovation, methodological innovation, and result innovation. No explanations are needed for the extracted results.
Input	Introduction of an article.
Output	**theoretical innovation**: [extracted theoretical innovation statement](if none, leave blank); **methodological innovation**: [extracted methodological innovation statement](if none, leave blank); **result innovation**: [extracted result innovation statement](if none, leave blank)"

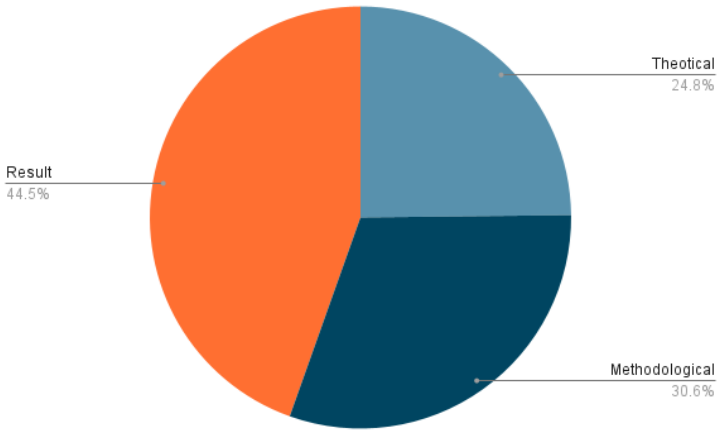


Figure 2. The proportion of three types of innovation evaluation sentences.

After completing the extraction and classification of peer review comments, we calculated the innovation evaluation scores given by peer reviewers to the papers using the extracted innovation evaluations. Specifically, we conducted a survey of common linguistic forms of innovation evaluations in peer review comments and constructed a lexicon of positive evaluations and a lexicon of negative innovation evaluations, as detailed in Tables A5 and A6 in the Appendix. The positive innovation lexicon includes sentiment words such as 'highly' and 'important', indicating that reviewers highly affirm the paper's innovativeness; whereas the negative innovation evaluation lexicon includes negative sentiment words such as 'lack' and 'insufficient', suggesting that reviewers find some aspect of the paper lacking in innovation. If an innovation evaluation sentence contains a positive innovation word, it is scored as 2 points; if it contains a negative innovation evaluation, it is scored as -1 point; all other ordinary innovation evaluation sentences are scored as 1 point. We accumulated the scores of each type of innovation evaluation sentence for each paper to obtain the innovation evaluation score for each paper's peer review, and then normalized it by dividing by the number of reviewers for each paper.

Calculating the novelty of academic papers

Novelty evaluation is vital for the promotion and management of innovation(Zhao & Zhang, 2025). To assist authors of academic papers in better promoting the innovative contributions of their research, we integrated existing metrics for measuring the innovativeness of academic papers to provide recommendations for authors on how to write about their innovative contributions. This study employed

the Novelty_U metric proposed by Uzzi et al., (2013) to measure the innovativeness of academic papers. This method is based on Schumpeter's (Chen et al., 2024) theory of combinatorial innovation, interpreting the innovation of a paper as a new combination of knowledge units and using the references of academic papers as a proxy for their knowledge sources to calculate the paper's novelty. The combination of knowledge, especially the combination of different types of knowledge, often produces novel knowledge(Chen et al., 2024). The advantage of this calculation method is that it can immediately determine the innovativeness of a paper after its completion, allowing authors to have a preliminary estimate of their research's novelty.

Specifically, this novelty calculation method quantifies the degree of innovation of an article by using the atypicality of the journal combinations to which the references of the academic paper belong. Firstly, the journals to which the references in each paper belong are paired two by two, and the frequency of occurrence of each journal pair is counted. Then, the references from the same year are recombined, ensuring that the length and temporal distribution of the references for each paper remain unchanged. This step is repeated multiple times to obtain the frequency of occurrence of each reference pair under random conditions. Finally, the atypicality z-score is calculated based on the actual occurrence and the frequency of occurrence under random conditions for each journal pair.

$$z\text{-}score = (obs_{ij} - rand_{ij})/std_{ij} \quad (1)$$

$$Novelty\_U = - P_{10}(z\text{-}score) \quad (2)$$

In this context, obs_{ij} represents the actual frequency of occurrence of two journals in a paper, $rand_{ij}$ is the expected frequency of occurrence of the two journals in a paper, and std_{ij} is the standard deviation of the occurrence of the two journals in a paper. Following the approach of Uzzi et al.(2013), we define the innovation of a paper as the negative value of the 10th percentile (P_{10}) of the z-scores of the reference journal pairs in the paper, sorted from smallest to largest. This means that the larger the value of Novelty_U, the more innovative the paper is considered to be.

Assessment of the intensity of promotional language

To promote transparency in academic communication and ensure that research findings are presented in a truthful and reasonable manner, we evaluated the intensity of promotional language extracted from the introductions of academic papers. For the intensity of promotional language in the introduction, we referred to the work of Sun et al.(2024) and designed two metrics. The first metric is the proportion of promotional language in the introduction of the paper (PL), which represents the

amount of effort the authors have spent on promotion in the introduction. The second metric is the proportion of promotional words in the introduction of the paper (PW). Promotional words refer to those with strong emotional connotations or exaggerated effects, typically used to attract attention, stimulate interest, or enhance the perceived value of something.

$$PL = Num(words\ of\ promotional\ language) / Num(words\ of\ Introduction) \times 100\% \quad (3)$$

$$PW = Num(promotional\ words) / Num(words\ of\ Introduction) \times 100\% \quad (4)$$

In this study, we focused on innovation-related promotional words such as 'new', 'unique', 'revolutionary', etc., which emphasize the innovative value of research findings. When constructing the promotional word lexicon, we referred to the lexicon proposed by Millar et al., (2022) based on promotional language in NIH grant applications, which includes 139 scientific promotional words. We further constrained the lexicon by stipulating that words are only recognized as promotional if they appear in the promotional language we extracted, thereby reducing bias from the different meanings of words. Although the research corpus we used consists of the introduction sections of academic papers, the purpose of the promotional language is similar to that of grant applications, both aiming to promote their research to reviewers or readers. Therefore, based on Millar's lexicon, we manually surveyed the promotional language extracted using a large language model and added some commonly used promotional words in academic papers.

We used these two metrics as proxies for promotional intensity and, in conjunction with the reference-based innovation metric for academic papers, investigated what level of promotional intensity could garner more positive evaluations from reviewers at the same level of innovativeness. Combining the reviewers' innovation scores calculated in section 3.1, we compared the top 5% and bottom 5% of papers based on their Novelty_U scores to observe how the use of promotional language in papers with different levels of novelty affects innovation evaluations in peer reviews.

Result

In this section, we present the differences in innovation focus between paper authors and reviewers, and analyze the impact of promotional language in academic papers on the peer review process in conjunction with innovation metrics.

Innovation Focus

Here, we utilize the contribution description sentences extracted from academic papers and the innovation evaluations extracted from peer review comments to

investigate *RQ1*, which is whether paper authors and reviewers place more emphasis on theoretical innovation, methodological innovation, or result innovation in scientific research within the field.

Innovation Focus of author In this section, we conduct a statistical analysis to determine which aspects of innovation paper authors are more focused on. We perform a statistical analysis on the various types of contribution-promoting sentences extracted using the large model. The statistical results are shown in Table 5, with the proportion of papers containing result innovation being the highest at 87.19%. This indicates that paper authors place greater emphasis on the innovation of results and are more likely to directly promote the innovation of their research findings in the introduction section of their papers.

Table 5. Total number and proportion of three promotion types.

Promotion Type	Number	Proportion
Theoretical innovation	5817	37.80%
Methodological innovation	4619	30.01%
Result innovation	13418	87.19%

From the proportion of various types of innovation, we can see that in some papers, authors claim that their research encompasses multiple types of innovation. We have counted the number of such papers and found that 33.4% of the papers contain at least two or more types of innovation-related contribution promotions. Among these, the proportion of papers that include all three types of innovation is the highest, at 26.46%. This allows us to explain why result innovation accounts for the largest proportion, namely that when paper authors make theoretical or methodological innovations, they often accompany these with innovative results. Authors of these papers believe that when declaring their contributions, they are comprehensively promoting the contributions of their research in all aspects. However, the majority of authors only promote the innovative results of their research in the introduction section, considering rseult innovation to be the core value of their study.

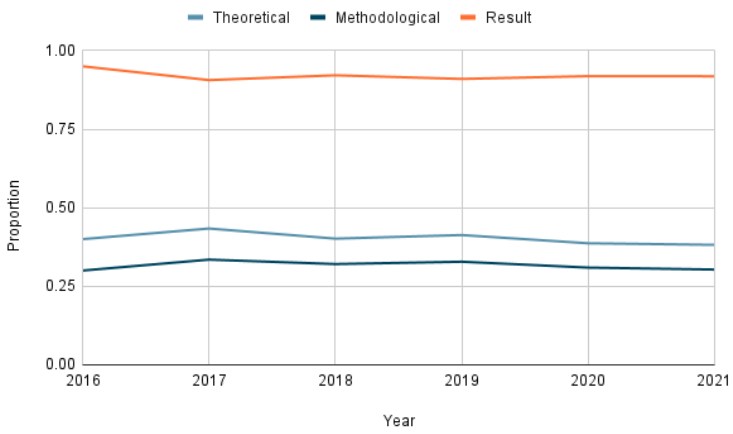


Figure 3. The changing innovation focus of authors over time.

We have examined the trend in the emphasis paper authors place on promoting various aspects of innovation by incorporating the publication dates of the papers. As shown in Figure 3, we can observe that the proportion of these three types of contribution promotions has remained relatively stable in recent academic writing. Most authors declare the innovative results of their research in the introduction section.

Next, we investigated which aspects of innovation are more emphasized by authors in different fields based on the domain of the papers. The results, as shown in Figure 4, reveal that authors across all five fields in Nature Communications place greater emphasis on the innovation of results, and their focus on theoretical innovation is also slightly higher than that on methodological innovation.

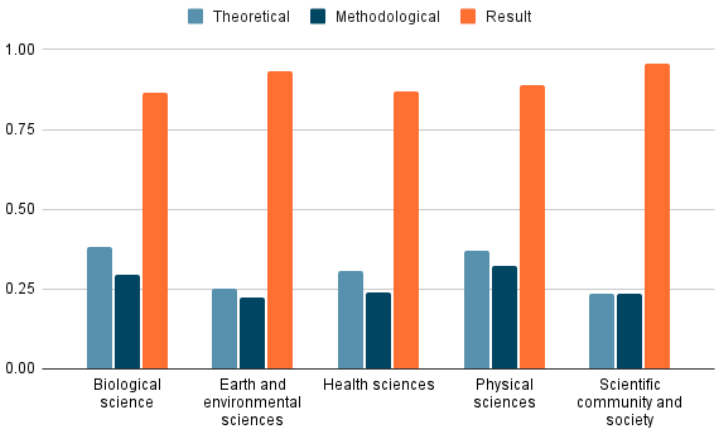


Figure 4. The innovation focus of authors in five different fields.

After extracting and statistically analyzing the contribution-promoting sentences in academic papers and the innovation evaluation sentences in peer review comments, we found that both reviewers and authors of academic papers are more inclined to evaluate and promote the innovative results of the papers. Analyzing the publication dates of the papers, we observed that the focus on innovation in the writing and review process of academic papers has remained relatively stable in recent years. When dividing the papers into different disciplinary fields for study, we discovered that authors from various disciplines have similar perspectives on the angles of contribution promotion, while reviewers' focuses differ significantly. For example, in the fields of scientific community and society and physical science, reviewers pay more attention to methodological innovation than in other fields.

Innovation Focus of reviewer We observed the focus of reviewers during the peer review process. From the peer review comments of 15,328 academic papers, we extracted 38,561 innovation evaluation sentences. After excluding sentences that evaluated the overall innovation of the papers, we categorized these sentences into theoretical innovation, methodological innovation, and result innovation. As shown in Figure 2, among all the extracted innovation evaluation sentences, result innovation accounted for the largest proportion at 50%, while theoretical innovation and methodological innovation evaluation sentences accounted for 22.4% and 27.6%, respectively. This indicates that reviewers place greater emphasis on the innovation of experimental results, such as new discoveries and conclusions in the research, when evaluating papers.

Analyzing the overlap in the evaluation of review comments, we found that 36.21% of the peer review comments for papers contained two or more types of innovation evaluations. Among these, 1,611 papers had all three types of innovation mentioned—theoretical, methodological, and result; 1,573 papers had both theoretical and result innovations mentioned; another 1,573 papers had both methodological and result innovations mentioned; and 686 papers had both theoretical and result innovations mentioned. This shows that the innovations in a paper often do not exist in isolation; for example, a paper may develop a new method and use it to discover new results. Reviewers tend to consider all aspects comprehensively when evaluating the innovation of a paper, which also explains why result innovation accounts for the largest proportion in innovation evaluations. This is because result innovation is more closely related to theoretical and methodological innovations, and when reviewers identify theoretical or methodological innovations, they are inclined to simultaneously evaluate the innovativeness of the results.

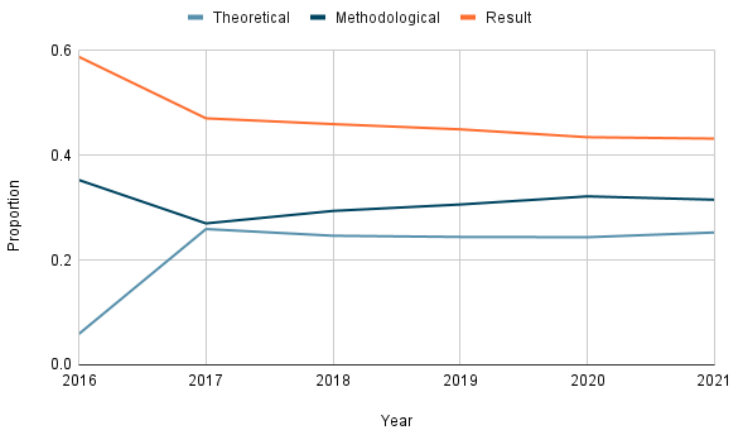


Figure 5. The changing innovation focus of reviewers over time.

We examined the proportion changes of different types of innovation evaluation sentences over time by incorporating the publication dates of the papers. Figure 5 illustrates the proportion changes of various types of innovation evaluation sentences from 2016 to 2021. Overall, the proportions of the three types of innovation evaluation sentences have remained relatively stable, with the proportion of methodological innovation evaluations showing an upward trend, while the proportion of result innovation evaluations has been declining. This indicates that in recent years, reviewers have been placing increasing emphasis on the innovation of research methods.

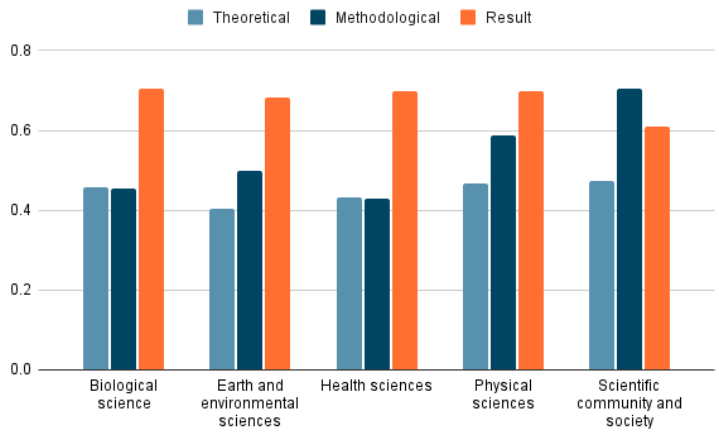


Figure 6. The innovation focus of reviewers in five different fields.

Next, we divided the study into five different disciplinary fields to examine the focus

of reviewers on innovation in peer review comments. The results, as shown in Figure 6, indicate that all disciplinary fields comprehensively review various aspects of innovation in papers, especially the innovation of results. In the field of scientific community and society, reviewers' attention to methodological innovation is particularly prominent, with 70.52% of the innovation evaluation sentences in the peer review comments for papers in this field being assessments of methodological innovation. This is because this field is interdisciplinary, requiring timely follow-up and integration of new methods from various disciplines to address current social issues; whereas the other four fields mostly rely more on the specialized knowledge and techniques of their respective fields to drive the production of more innovative research outcomes.

Differences in Innovation Focus Between Authors and Reviewers

In section 4.1, we analyzed which aspects of innovation are focused on in academic papers and peer review comments, respectively. It can be observed that although both paper authors and reviewers pay considerable attention to the innovation of paper results, their focuses still differ. For instance, reviewers tend to be more comprehensive when examining papers, also paying attention to the theoretical and methodological innovations of the papers. Therefore, in this section, we address *RQ2*, which is what differences exist between authors and reviewers when evaluating the innovative points of a paper, and which innovations mentioned by authors in the introduction are recognized by reviewers?

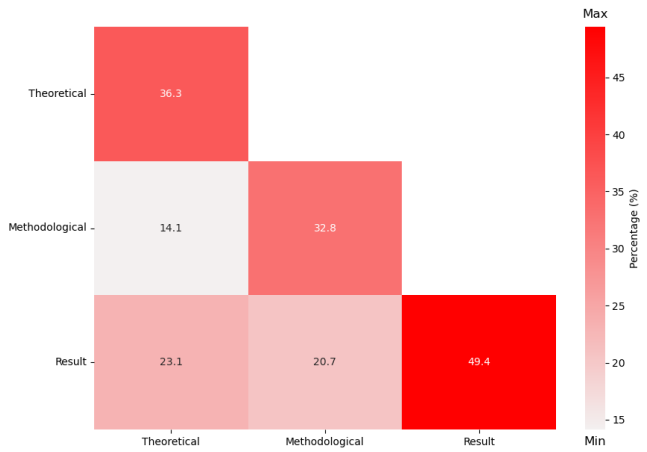


Figure 7. Heatmap of overlap ratio between reviewers' positive evaluations and authors' promotional language.

We investigated the relationship between the contribution-promoting sentences provided by authors and the innovation evaluations given by reviewers in the same academic paper, and used this to create a heat map. As shown in Figure 7, the horizontal axis represents the types of innovation described by the paper authors, and the vertical axis represents the types of positive innovation evaluations given by peer review experts. The content of each cell indicates the proportion of papers that received corresponding positive evaluations from review experts when authors provided that type of contribution-promoting sentence. For example, 49.4% of the papers that promoted their research result innovations were recognized by peer review experts; 20.7% of the papers that promoted their research methodological innovations also received positive evaluations for theoretical innovation from the review experts. From the figure, we can see that when describing research result innovations in a paper, it is more likely to receive positive evaluations from review experts compared to the other two types of innovation. Considering that a paper may contain multiple innovations and that review experts may also make innovation evaluations on various aspects of the paper, we can observe that theoretical innovation is closely linked with result innovation, with 23.1% of the papers proposing theoretical innovation and receiving positive evaluations from reviewers for their result innovations.

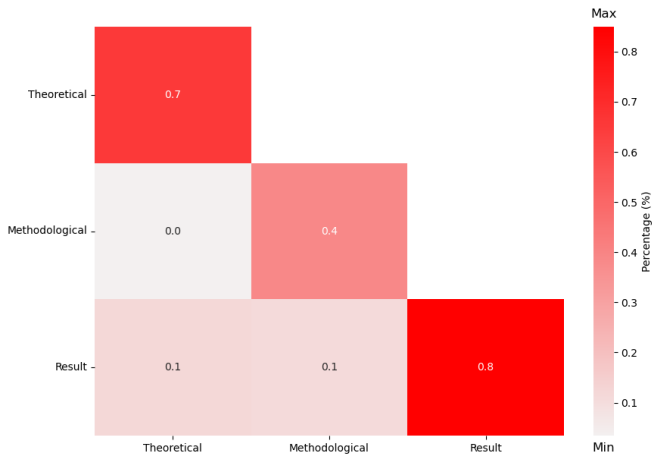


Figure 8. Heatmap of overlap ratio between reviewers' negative evaluations and author promotional language.

However, during the review process, reviewers may also provide negative

evaluations, such as considering that the paper is not as novel as claimed, or that similar topics have been studied before but the authors did not mention them. We have also conducted statistics on these papers, and the results are shown in Figure 8. Since the data we used only includes the original texts of accepted papers and peer review corpora, the quality is generally high, and there are fewer negative evaluations in the peer review comments. Among them, negative evaluations mostly appear in the innovation points claimed by the authors themselves. For example, 0.8% of the authors promoted the innovativeness of their results, but the reviewers considered their results not to be innovative.

Through the above research, we can find that the innovation promotion in the introduction of a paper does indeed draw the attention of reviewers to that type of innovation. However, if the contribution is misrepresented or improperly promoted, it is more likely to be refuted by review experts. As for innovations not declared in the introduction, review experts rarely give negative opinions.

The Relationship Between Promotional Intensity and Peer Review

To delve deeper into the impact of promotional language in the introduction of papers on the review process, we address *RQ3* in this section: Does promotional language influence review comments, and what level of promotional intensity is appropriate in a paper? To tackle this issue, we incorporated the novelty calculation metric Novelty_U based on references proposed by Uzzi et al.(2013). The advantage of this metric is that it can be calculated immediately upon completion of the writing, allowing for an assessment of its novelty. To measure the intensity of promotional language in the introduction of papers, we use the proportion of promotional language in the introduction to gauge the effort authors spend on promoting innovation, and the proportion of promotional words in the introduction to assess the degree of promotion.

Since the inherent innovativeness of a paper is a key factor influencing the scores given by peer review experts, we controlled for the paper's own innovativeness to study how the promotional language in a paper affects reviewers' comments when the paper's innovativeness is the same or similar. We identified the top 5% most innovative papers and the least innovative 5% of papers based on Novelty_U from all the papers to investigate the relationship between the promotional language in their introductions and the innovation evaluations provided by peer review.

Table 6. Correlation between the proportion of promotional language in the introduction and peer review score in top 5% Novelty_U.

Variable	promotional language	promotional word	promotional language square	promotional word square	peer review score
promotional language	1.000	.608***	1.000***	.608***	0.044
promotional word	.608***	1.000	.608***	1.000***	.089*
promotional language square	1.000***	.608***	1.000	.608***	0.044
promotional word square	.608***	1.000***	.608**	1.000	.089*
peer review score	0.044*	.089**	0.044*	.089**	1.000

Note: *: p <0.05, **: p < 0.01, ***: p < 0.001.

Table 7. Correlation between the proportion of promotional language in the introduction and peer review score in bottom 5% Novelty_U.

Variable	promotional language	promotional word	promotional language square	promotional word square	peer review score
promotional language	1.000	.850***	.483***	.850***	0.019
promotional word	.850***	1.000	.518***	1.000***	-0.021
promotional language square	.483***	.518***	1.000	.518***	-0.001
promotional word square	.850***	1.000***	.518***	1.000	-0.021
peer review score	0.019	-0.021*	-0.001	-0.021*	1.000

Note: *: p <0.05, ***: p < 0.001.

We first conducted a study on the most innovative portion of the papers. Table 6 shows the relationship between the proportion of promotional sentences in the introduction, the proportion of promotional words, and the peer review scores. We

can see that both the proportion of promotional language and the proportion of promotional words are positively correlated with peer review scores, with the correlation for promotional words being significant. This indicates that for the most innovative introductions, the more promotional words used, the more likely it is to receive positive evaluations from review experts.

For the least innovative 5% of papers, as shown in Table 7, we find that the promotional words in their introductions are negatively correlated with peer review scores. This means that for papers lacking in innovation, it is not advisable to excessively promote their innovativeness in the introduction, as it may cause dissatisfaction among reviewers.

In summary, we have found that the use of promotional words in papers is related to their own innovativeness. When a paper possesses strong innovativeness, more promotional words and language can be used in the introduction to better convey the novelty of the research to review experts, garnering more positive evaluations and facilitating the publication of the paper. However, when a paper lacks innovativeness, it should avoid exaggerate in the introduction to prevent causing aversion among reviewers.

Discussion

Implications

We will elaborate on the implications of this study from both theoretical and practical perspectives.

Theoretical implications: This study explores the cognitive gap between authors and peer reviewers regarding the perception of novelty in academic papers, specifically using *Nature Communications* as a case study. Despite the availability of open peer review datasets, these resources often lack standardized writing conventions, which complicates the extraction of meaningful insights.

We developed a novel "rules + machine learning" approach to effectively extract novelty assessment sentences from peer review comments, demonstrating improved accuracy in identifying relevant evaluations. Furthermore, we utilized the large language model DeepSeek to automatically extract and categorize contribution statements from the introductions of academic papers. This innovative method transcends traditional rule-based information extraction, enabling a more nuanced understanding of how novelty is communicated in academic writing.

Our findings reveal significant discrepancies in how authors and reviewers perceive the novelty of research contributions, as evidenced by our analysis of the extracted data. By integrating literature-based novelty measurement indicators, this study not

only provides a new framework for examining peer review comments but also highlights the potential for further research into the communication of innovation in various academic contexts.

Overall, this research contributes to the theoretical discourse on peer review practices by offering a systematic approach to assess and understand the dynamics of novelty evaluation, paving the way for future studies to explore other dimensions of author-reviewer interactions.

Practical implications: The experimental conclusions of this study can provide recommendations for researchers in academic paper writing. To better enable review experts to understand the innovative aspects of the paper, authors should articulate the corresponding points of innovation in the introduction section, but also avoid over-promotion. This is because while promotional language in the introduction can benefit genuinely innovative parts, improper promotion may also raise doubts among review experts.

After completing the writing of their papers, authors can refer to existing academic paper innovation metrics to estimate the novelty of their own research. If the research is highly novel, more promotional language and words can be used in the introduction to make reviewers more aware of the innovative aspects of the research; if the novelty is low, the use of aggressive promotional words should be avoided to prevent exaggerate from raising doubts among reviewers.

Limitations

Indeed, this study has certain limitations. Firstly, the effectiveness of extracting contribution-promoting sentences from the original academic papers and innovation evaluation sentences from peer review comments needs improvement. The accuracy of the extraction results may affect the validity of subsequent conclusions. Particularly, the lack of uniform standards in peer review comments, the varying language styles of reviewers, and some implicit evaluations of paper innovation have impacted our extraction accuracy to some extent. Secondly, the corpus we used is limited to papers published in *Nature Communications*. Although this journal covers multiple disciplines within the natural sciences, the findings of this study have not been fully validated in some disciplinary fields. Moreover, different journals may have different review requirements, and the focus of reviewers on innovation may change with alterations in review criteria. However, currently, only a minority of scientific publications choose to open peer review comments, so this study has only made a preliminary exploration of the research question using papers published in *Nature Communications*. Lastly, the dataset used in this study consists solely of

accepted papers and their peer review comments, and does not include data from rejected papers, which may also affect the generalizability of the study's results.

Conclusion and future works

In this paper, we investigated the cognitive differences between paper authors and reviewers regarding the innovation of papers during the writing and review process, provided recommendations for academic paper writing, promoted appropriate promotion in the academic paper writing process, and reduced the cognitive gap between paper authors and review experts.

In future work, firstly, we aim to improve the accuracy of extracting contribution-promoting sentences from the original academic papers and innovation evaluation sentences from peer review comments. Currently, large models have shown superior performance in natural language processing tasks, and we can attempt to use different large models to optimize this extraction task. Optimizing this extraction task can not only enhance the understanding of innovation-related writing in academic papers but also extend to various aspects of knowledge in academic papers, such as the extraction and analysis of future work sentences, thereby advancing the development of information extraction in academic papers. Secondly, the dataset for this study can be expanded to analyze the original texts and peer review comments of rejected papers, investigating whether the reasons for rejection are related to over-promotion or insufficient promotion, to supplement and extend the conclusions of this study. Finally, we can incorporate disciplinary fields into the study of this issue to observe whether the focus of paper authors and review experts varies across different disciplinary fields.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Grant No.72074113).

References

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text* (arXiv:1903.10676). arXiv. https://doi.org/10.48550/arXiv.1903.10676

Chen, Z., Zhang, C., Zhang, H., Zhao, Y., Yang, C., & Yang, Y. (2024). Exploring the relationship between team institutional composition and novelty in academic papers based on fine-grained knowledge entities. *The Electronic Library*. https://doi.org/10.1108/EL-03-2024-0070

Costello, A., Fedorova, E., Jin, Z., & Mihalcea, R. (2023). *Editing a Woman's Voice*

(arXiv:2212.02581). arXiv. https://doi.org/10.48550/arXiv.2212.02581

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., … Pan, Z. (2024). *DeepSeek-V3 Technical Report* (arXiv:2412.19437). arXiv. https://doi.org/10.48550/arXiv.2412.19437

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, *17*(1), e0259238. https://doi.org/10.1371/journal.pone.0259238

Jiang, G. K., & Jiang, Y. (2023). More diversity, more complexity, but more flexibility: Research article titles in TESOL Quarterly, 1967–2022. *Scientometrics*, *128*(7), 3959–3980. https://doi.org/10.1007/s11192-023-04738-x

Leahey, E., Lee, J., & Funk, R. J. (n.d.). What Types of Novelty Are Most Disruptive? *American Sociological Review*.

Li, Z. (2022). Is academic writing less passivized? Corpus-based evidence from research article abstracts in applied linguistics over the past three decades (1990–2019). *Scientometrics*, *127*(10), 5773–5792. https://doi.org/10.1007/s11192-022-04498-0

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., & Zhang, C. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology*, *70*(5), 462–475. https://doi.org/10.1002/asi.24126

Millar, N., Batalo, B., & Budgell, B. (2022). Trends in the Use of Promotional Language (Hype) in Abstracts of Successful National Institutes of Health Grant Applications, 1985-2020. *JAMA Network Open*, *5*(8), e2228676. https://doi.org/10.1001/jamanetworkopen.2022.28676

Parker, T. H., Griffith, S. C., Bronstein, J. L., Fidler, F., Foster, S., Fraser, H., Forstmeier, W., Gurevitch, J., Koricheva, J., Seppelt, R., Tingley, M. W., & Nakagawa, S. (2018). Empowering peer reviewers with a checklist to improve transparency. *Nature Ecology & Evolution*, *2*(6), 929–935. https://doi.org/10.1038/s41559-018-0545-z

Pearson, W. S. (2020). Research article titles in written feedback on English as a second language writing. *Scientometrics*, *123*(2), 997–1019. https://doi.org/10.1007/s11192-020-03388-7

Peng, H., Qiu, H. S., Fosse, H. B., & Uzzi, B. (2024). Promotional language and the adoption of innovative ideas in science. *Proceedings of the National Academy of Sciences*, *121*(25), e2320066121. https://doi.org/10.1073/pnas.2320066121

Sagi, I., & Yechiam, E. (2008). Amusing titles in scientific journals and article citation. *Journal of Information Science*, *34*(5), 680–687. https://doi.org/10.1177/0165551507086261

Song, N., Chen, K., & Zhao, Y. (2023). Understanding writing styles of scientific papers in the IS-LS domain: Evidence from abstracts over the past three decades. *Journal of Informetrics*, *17*(1), 101377. https://doi.org/10.1016/j.joi.2023.101377

Stelmakh, I., Shah, N., & Singh, A. (n.d.). *On Testing for Biases in Peer Review*.

Sun, Z., Cao, C. C., Liu, S., Li, Y., & Ma, C. (2024). Behavioral consequences of second-person pronouns in written communications between authors and reviewers of scientific papers. *Nature Communications*, *15*(1), 152. https://doi.org/10.1038/s41467-023-44515-1

Teplitskiy, M., Peng, H., Blasco, A., & Lakhani, K. R. (2022). Is novel research worth doing? Evidence from peer review at 49 journals. *Proceedings of the National Academy of Sciences*, *119*(47), e2118046119. https://doi.org/10.1073/pnas.2118046119

Transparent peer review for all. (2022). *Nature Communications*, *13*(1), 6173, s41467-022-33056–33058. https://doi.org/10.1038/s41467-022-33056-8

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474

Wheeler, M. A., Vylomova, E., McGrath, M. J., & Haslam, N. (2021). More confident, less formal: Stylistic changes in academic psychology writing from 1970 to 2016. *Scientometrics*, *126*(12), 9603–9612. https://doi.org/10.1007/s11192-021-04166-9

Wicherts, J. M. (2016). Peer Review Quality and Transparency of the Peer-Review Process in Open Access and Subscription Journals. *PLOS ONE*, *11*(1), e0147913. https://doi.org/10.1371/journal.pone.0147913

Wu, W., Zhang, C., Bao, T., & Zhao, Y. (2025). SC4ANM: Identifying optimal section combinations for automated novelty prediction in academic papers. *Expert Systems with Applications*, *273*, 126778. https://doi.org/10.1016/j.eswa.2025.126778

Zhao, Y., & Zhang, C. (2025). A review on the novelty measurements of academic papers. *Scientometrics*, *130*(2), 727–753. https://doi.org/10.1007/s11192-025-05234-0

Appendix: Dictionaries and Rules for Coding

Table A1. Innovation Signifying Terms Dictionary.

New			
novel	new	innovative	creative
novelty	uncover	fill the knowledge gap	methodological step
breakthrough	groundbreaking	pioneering	trailblazing
disruptive	revolutionary	unprecedented	advancement
introduce	propose	unique	originaL

When these words appear after the word “new” or ‘original’, the word “new” or ‘original’ is not tagged as NEW: 'figure', 'we', 'our', 'table', 'version', 'fig', 'review', 'paragraph', 'claim', 'manuscript', 'comment', ' added', 'new text', 'sample', 'avoid', 'supp'.

We exclude mentions of “first” as NEW, unless one of these words appears immediately afterward: principle', 'result', 'observation', 'attempt', 'experiment', 'synthesis', 'study', 'comprehensive', 'application', 'description', 'describe', 'evidence', 'design', 'time'.

Table A2. Theoretical Innovation Terms Dictionary.

Theoretical			
concept	generalize	mechanism	synthesize
theoretical	explanation	hypothesis	model
term	theory	insight	point
idea	hypotheses	thesis	explain

Table A3. Methodological Innovation Terms Dictionary.

Methodological			
method	analysis	classification	experiment
methodology	strategy	analyze	criteria
formula	procedure	technique	apparatus
criterion	index	process	technology
design	means	protocol	tool
equipment	measure	quantify	calculate
test	examine	experimental	approach

Table A4. Result Innovation Terms Dictionary.

Result			
structure	confirm	finding	observation
response	correlation	found	outcome
result	effect	discovery	prove
support	evidence	identify	rate
show	observation	data	report

If a sentence contains two or more different terms from the innovation aspect dictionaries, we conduct a syntactic analysis to observe which terms from Table A1 modify the terms from Table A2 to A4, thereby determining the innovation type of the sentence.

Table A5. Positive Innovation Terms Dictionary.

Positive			
highly	interest	completely	important
strong	indeed	surprise	extensive
sound	timely	extremly	incredibly
remarkably	exceptionally	significantly	thorough
thoughtful	clever	unquestionably	robust

Table A6. Negative Innovation Terms Dictionary.

Negative			
lack	insufficient	inaccurate	ineffective
unconvincing	inappeoriate	slight	

When the word “not, “no,” or “lack” appears near the word “new,” it is tagged as negative.

https://doi.org/10.51408/issi2025_018

Examining the Patenting Activities of Universities in the Middle East and North Africa

Jamal El-Ouahi

j.el.ouahi@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, Leiden (Netherlands)
Clarivate Analytics, Dubai Internet City, Dubai (United Arab Emirates)

Abstract

This paper aims to examine universities' patenting activities in the Middle East and North Africa (MENA) region. Patent data from Derwent Innovation is analyzed to provide key insights about such activities. Saudi Arabia leads the region in terms of the number of patents, followed by Turkey and Morocco. These three countries, which represent 87% of all academic patents, are also home to the most patenting academic institutions. Although the academic sector in MENA grew its patenting activity faster than the world, its patent volume base is still relatively low. The results also show the profile of the technological developments covered in MENA academic patents. Some of these inventions directly tackle societal health-related issues but also public environmental ones. The main academic assignees show a certain degree of collaboration with academic and corporate organizations. This study provides important input to research managers as well as policymakers to assess the research produced by universities from a technological and economic perspective.

Introduction

For the past 20 years, research-intensive universities have been increasingly subject to quantitative research evaluation with various expectations to contribute more to societal and economic development (Clark, 1998; Mejlgaard & Ryan, 2017). At the same time, numerous calls have been made to reform research evaluation and move from quantitative to more inclusive and qualitative assessment. For example, Wilsdon et al. (2015) argue that evaluation should promote the diversity and plurality of research in *The Metric Tide* report. In Europe, 350 institutions, including research organizations, funding agencies and assessment groups have recently pledged to sign such a reform call (Directorate-General for Research and Innovation, 2022). This call to reform research assessment encompasses multiple dimensions such as the recognition of various contributions that researchers make to both science and society. Such contributions extend beyond traditional journal publications and include diverse scientific outputs. This study addresses this issue from the perspective of assessing the economic impact of research produced by Universities.

Historical models of research and innovation have traditionally described a uni-directional flow of funding and knowledge between government, academia, and industry (Pavitt & Walker, 1976). Later, Limoges et al. (1994) introduced the Mode 2 knowledge production framework, which represents a more collaborative and interdisciplinary approach to knowledge production. Mode 2 is characterized

by the integration of different knowledge systems, including academic and non-academic perspectives, and emphasizes the co-production of knowledge by multiple stakeholders, including researchers, industry partners, and policymakers. Mode 2 research tends to be more applied and problem-oriented, with a focus on addressing real-world challenges. This framework can help to contextualize the knowledge dynamics of universities in the Middle East and North Africa, where there is often a tension between the traditional academic knowledge production and the demand for practical, socially relevant knowledge (Altbach, 2009; Hanafi & Arvanitis, 2015).

The economic impact of scientific research is a component of its societal impact. It is widely acknowledged that technological innovation has a significant role in the economic growth and competitiveness of institutions, regions, and countries (Tödtling & Trippl, 2005). The two most popular indirect measures of innovation are R&D expenditures, which serve as an indicator of the process' input, and patent data, which serves as a measure of inventive activity's output (Basberg, 1987). Patents are mainly used due to the large amount of information available across borders and regions. Also, in the context of a knowledge-intensive economy, patents are a crucial tool in the protection of intellectual property.

There is a massive literature on innovation activities in the academic sector (Dornbusch et al., 2013; Lissoni, 2013; Perkmann et al., 2013; van Zeebroeck et al., 2008). This literature covers mostly Western countries. However, literature on patenting activities by universities in emerging nations such as in the Middle East and North Africa (MENA) region is rather scarce. Only a few studies covered the patenting activity by the academic sector in North Africa (Landini et al., 2015), in Iran (Noruzi & Abdekhoda, 2012) and Turkey (Uzun, 2001). In this paper, I attempt to address this gap by examining certain aspects of the innovation activities of universities in this specific region in recent years. Although innovation studies go beyond patentometrics, various insights can be gained by examining the data of patent documents. Indeed, patents constitute a rich source of data from technology and scientific research perspectives. This quantitative and empirical study explores the patenting activities of research universities in MENA. Based on this topic, the following general hypothesis is proposed to investigate the knowledge dynamics involved in creating and transferring knowledge within the MENA region:

Hypothesis: The Mode 2 framework of university-industry collaboration is positively associated with the patenting activities of universities in the Middle East and North Africa.

Specifically, in this empirical study, I address the following research questions:

- What are the recent trends of technological advancements developed by research universities in the Middle Eastern and North African nations from a patent's perspective?
- What are the technological characteristics of such developments?

- To which extent does academia collaborate with the industry in MENA in terms of patenting activity?

These aspects provide insights into the contribution of research universities to societal impact from a patent's lens and support a country's future development. Such insights are also particularly helpful for research assessment and decision-making when formulating science and technology policies. This study is organized as follows. The next section describes the data used to analyze the patenting activities by the academic sector in MENA. Then, the findings are presented in the following section. Finally, the results of this study are discussed in the last section of this paper.

Methods and data

Data source

The patent collection used for this study was developed by using the full patent content on Derwent Innovation, provided by Clarivate. Derwent Innovation includes the Derwent World Patent Index (DWPI), which covers over 59 patent authorities worldwide and 2 journal sources. DWPI provides curated data including editorially enhanced titles and abstracts in the English language.

Data counting definition

The "patent families" are the building blocks of the DWPI database. As soon as it is published, each associated patent application and granted patent is added to the related DWPI family record. As a result, rather than referring to specific patent documents, all counts of records in this analysis refer to patent families or inventions. For instance, unless otherwise stated, all analyses in this study will count, for example, a combined United States patent application and European patent application as a single innovation family or one innovation. This gives a more accurate image of the overall level of innovation in a specific field as well as a more accurate measure of the level of inventive activity from a particular organization within the corresponding technological domain. Entity names for patents were cleaned and harmonized, to the greatest possible extent. Known subsidiaries and merger and acquisition entities were consolidated under a single company name for a more realistic view of the collaborating corporations. Also, in terms of co-patenting, a full counting approach is used in this study.

Geographic coverage

The following nations make up the MENA region, according to the World Bank (2019): Algeria, Bahrain, Djibouti, Egypt, Iran, Iraq, Jordan, Kuwait, Lebanon, Libya, Morocco, Oman, Palestine, Qatar, Saudi Arabia (KSA), Syria, Tunisia, the United Arab Emirates (UAE) and Yemen. In this study, Pakistan, Afghanistan and

Turkey are also considered as commonly included in the MENA region (MENAP and MENAT).

Search string creation and quality control

The search for relevant patents was conducted using the so-called 'expert search' of Derwent Innovation. The search string for the patent analysis was developed iteratively, with the search results being examined and assessed to guide and improve the search query's accuracy. Necessary changes are made to the keywords used for the assignee names of academic institutions. This procedure is repeated until only slight differences in the results are produced by revisions. The period covered in this study is 2008-2021. The final search query consists of a combination of various fields and is shown below:

PAOC=(AE or AF or BH or DJ or DZ or EG or IQ or IR or JO or KW or LB or LY or MA or OM or PK or PS or QA or SA or SY or TN or TR or YE) and PA=(univ or uni or inst* or acad*) and PY > (2007) and PY < (2022);*

- PAOC represents the country code of the patent assignee/applicant
- PA is the assignee or applicant name
- PY stands for Publication Year

The dataset under study consists of 18,348 individual patents, classified as 10,010 individual DWPI invention families.

Visualising patents landscapes with ThemeScape

ThemeScape is a text-mining application that analyzes text sources (Clarivate, 2022). Its algorithms do not require a thesaurus or other external sources of information. After analyzing the text in multiple documents, it groups together the documents that share related text and separates the documents with less related text. The result of such analysis is presented as a topographical map. Each document is placed on the map in a unique position that is the vector sum of its relatedness to all the other documents.

ThemeScape uses the frequency of occurrence and co-occurrence of words to select the topics of interest. Then, it aggregates words that have a common stem, but it does not directly aggregate synonyms. Instead, synonyms may be clustered under a common theme because of the other words that co-occur with those synonyms. In other words, terms are identified as synonyms only by co-clustering based on common themes. For example, "battery" and "cell" may be grouped together because of the co-occurrence in the same documents of terms such as "electrode" or "rechargeable". On the other hand, "battery" and "cell" may also be separated if the map contains a set of electric power and biology patents, where the term "cell" has different meanings.

The topographical maps presented by ThemeScape are built on a random selection of a first patent and sequential calculation of the relationships of all the other patents. The orientation of the map is randomly set, and the different directions have no significance. Only the proximity of points within the map is relevant, and co-clustered patents are highly likely to share common concepts.

Findings

Recent trends of patenting activities by research universities in MENA by country

Before reporting the trends of patenting activities by assignees affiliated with research universities in MENA, I analyzed their total patent output at the country level. This analysis is shown in Figure 1. Research institutions in Saudi Arabia lead the MENA region in terms of patent filings with 48% of the patents filed by the academic sector in the region. Turkey (28%) and Morocco (11%) follow. The academic institutions in these three countries cumulate 87% of all the patents under study. Also, several countries such as Algeria, Bahrain, Iraq, Libya, Palestine, Afghanistan, Syria, Yemen and Djibouti show a very low output, with less than 10 patents filled during the study period. These results suggest that research institutions in Saudi Arabia, Turkey, and Morocco have made strides in patent registration globally.

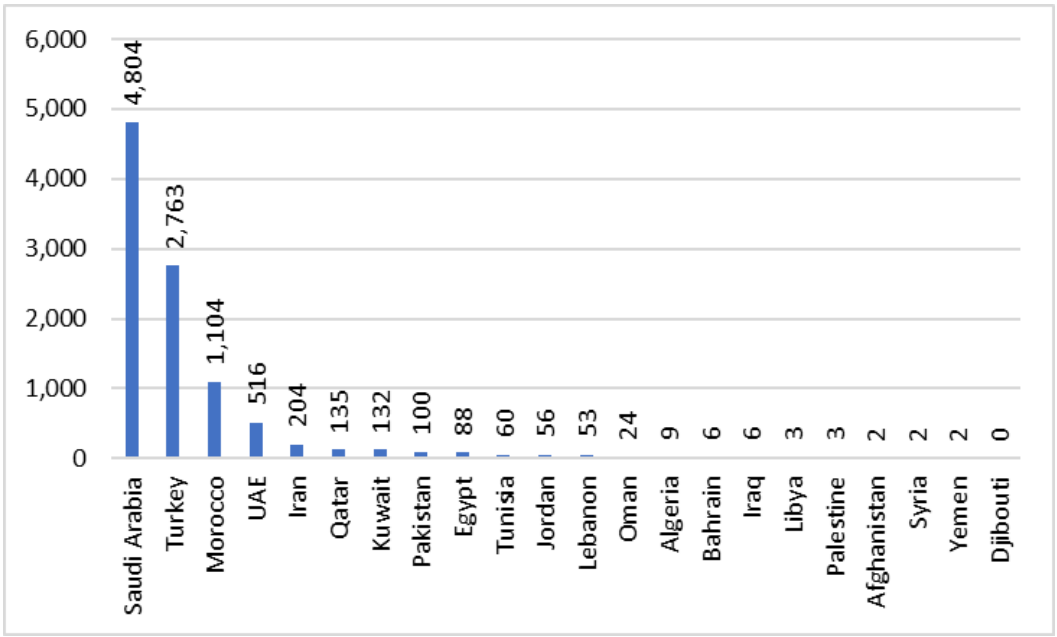


Figure 1. Number of patents published between 2008 and 2021 by assignees affiliated with research institutions in MENA.

The top 20 assignees within the dataset under study in terms of number of patents are shown in Figure 2. These most productive institutions are located in Saudi

Arabia (8), Turkey (6), Morocco (2), UAE (2), Qatar (1) and the US (1). The presence of the US suggests a certain level of international co-patenting activities by MENA universities with the US, specifically with The Massachusetts Institute of Technology (MIT) found in 72 patents as a co-assignee. Also, Saudi Arabian Oil Company (Aramco) co-patented 229 with at least one academic institution from MENA, which makes it the largest co-patenting corporate entity with Academia in MENA and more precisely with King Fahd University of Petroleum and Minerals (KFUPM). This evidence provides support to the hypothesis of this study.

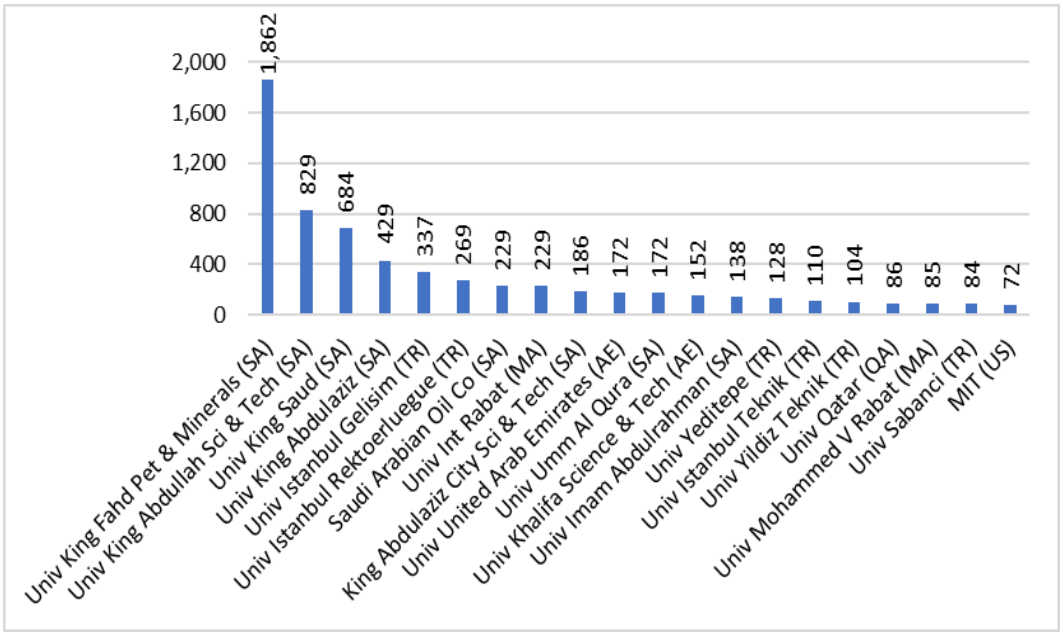


Figure 2. Top 20 Institutions by number of patents in the dataset.

Figure 3 shows the trends of patenting activities by the academic sector in MENA between 2008 and 2021 for countries with more than 200 patents (Saudi Arabia, Turkey, Morocco, UAE, and Iran). The number of patents grew from 46 in 2008 to 2,164 in 2021 for the whole region, representing a growth of 4,604%. Following the methodology explained earlier, the academic sector across the world published 16,040 patents in 2008 and 389,656 in 2021, which represents a growth of 2,329%. The patenting activity by the academic sector grew faster in MENA, although the MENA institutions started from a very low base in 2008 which explains in part this impressive increase.

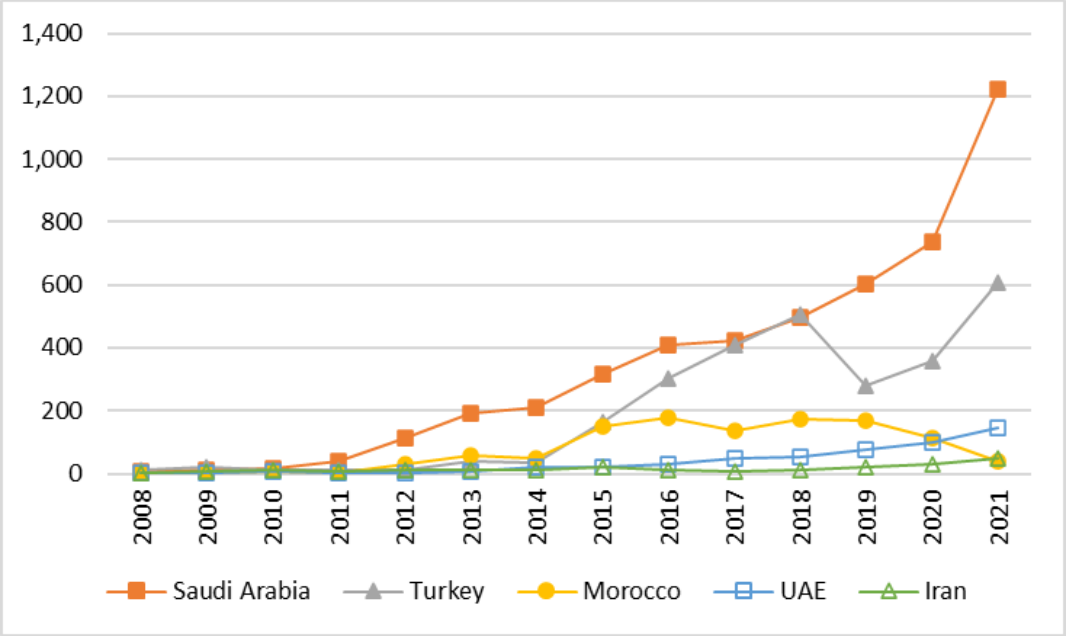


Figure 3. Trends of the number of patents published between 2008 and 2021 by research institutions in Saudi Arabia, Turkey, Morocco, UAE and Iran.

Patent filings by academic institutions in Saudi Arabia have gradually increased over the past few years. Saudi Arabia's remarkable output increase might be due to the effects of the kingdom's 'Vision 2030', the policies set locally, and initiatives led by the Saudi patent office. Saudi Arabia and Turkey had the same patent output level by academic institutions in 2018. However, Turkish research institutions saw a decrease in their patenting activity in 2019. Since then, academic institutions in Turkey and Saudi Arabia saw their output grow at the same rate. Moroccan institutions have initially shown growth in terms of the number of patents. Their output stabilized between 2015 and 2019 and then declined to reach the 2014 level. Academic institutions in the UAE have also experienced an increase in their number of patents since 2015. We notice a similar trend for research organizations in Iran.

A profile of patenting activities by Academia in MENA

In this sub-section, two aspects of the patenting activities are analyzed: their geographic distribution in terms of legal jurisdictions and then their technical coverage.

A patent application only provides a potential monopoly on the covered technology it covers within the legal jurisdiction of the issuing authority. As a result, applicants must submit patent applications to multiple patent bodies and jurisdictions in order to obtain broader geographic patent protection. The level and timeline of patent protection in the various patent jurisdictions are analyzed in Figure 4. The

authorities with more than 500 patents filed are shown individually, and the others are combined together into the 'Other' authority.

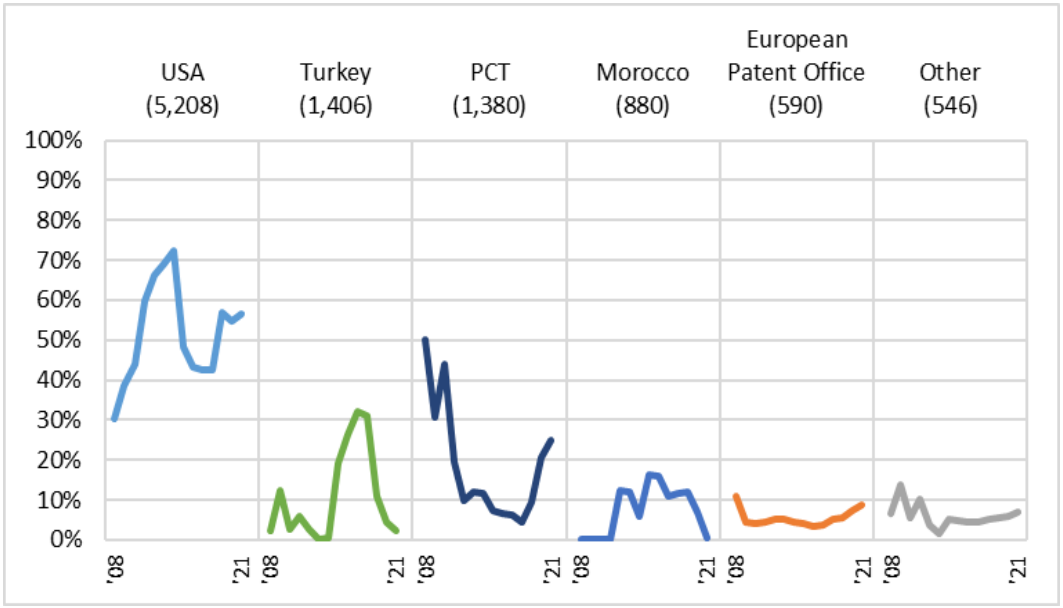


Figure 4. Share of inventions filed by patent authority and by assignees affiliated with research institutions in MENA between 2008 and 2021.

Patent protection continues to be most often sought in the United States, with filings in the US the predominant jurisdiction in the dataset under study. The academic institutions in MENA also commonly use the Patent Cooperation Treaty (PCT) application route, which provides a patent filing fast track for individual later patent applications in countries designated by the applicant. It is worth reminding that the PCT filings do not produce granted patents themselves. Indeed, patent prosecution must be still sought at individual patent authorities. On the one hand, the share of inventions at the PCT level initially decreased and then increased in the recent years. On the other hand, protection was also commonly sought at the Turkish and Moroccan Patent Office. These two authorities have seen sharp increases then declines in terms of share of inventions filed by academic institutions in MENA. Invention protection is also commonly sought at the European Patent Office. Such protection provides potential EPO member state-wide protection. Filings in the US, at the EPO and via the PCT application process are popular and recent. This is the usual protection regime within the European community, and it might suggest that MENA academic institutions collaborate with peer institutions from Europe. Second-tier application locations include China, South Korea, Saudi Arabia, Germany and Canada.

As for the technical focus of the patents dataset under study, the dataset was segmented into major research categories using the Derwent World Patents Index

(DWPI) patent classification scheme for categories with more than 100 inventions. This taxonomy is shown in Figure 5.

The largest technical fields include *Polymers & Plastics* (24%), *Pharmaceuticals* (19%) and *Computing & Control* (19%). The number of *Polymers & Plastics* patents increased from 2 patent filings in 2008 to 320 in 2021. *Pharmaceuticals* also saw a large increase in patenting activity with 14 patents in 2008 and 200 in 2021. Similarly, the number of *Computing and Control* patents increased from 2 patents in 2008 to 174 in 2021. It is worth reminding that there is a high level of overlap between some of the fields shown in Figure 5 such as *Food, Fermentation, Disinfectants, Detergents* and *General chemicals*, as patents with classifications pertinent to both fields have been categorized into multiple industrial fields.

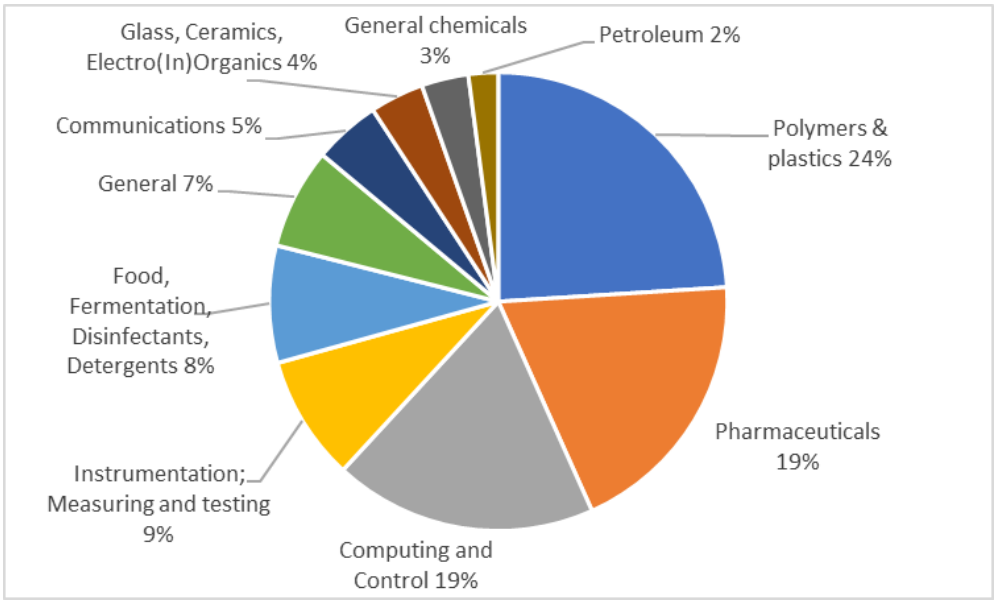


Figure 5. Number of inventions by technical area by assignees affiliated to research institutions in MENA between 2008 and 2021.

Next, the technical nature of the inventions of the dataset under study has been summarized using *ThemeScape* (Clarivate, 2022). Such visualization is shown in Figure 6 and provides the common themes and concepts within the dataset.

The contour lines on the map diminish in terms of circumference and are meant to encircle regions of higher document concentration. The density is also represented by the map colors. White snow-capped peaks represent the highest density, while blue areas indicate low density. The words included in the map are those shared by the patent documents in their DWPI abstracted form and have been selected by ThemeScape based on the term frequency. The individual dots on the map represent single patents. Dots are not shown for all the documents, and instead, represent a sampling that allows the other features of the map to be discerned.

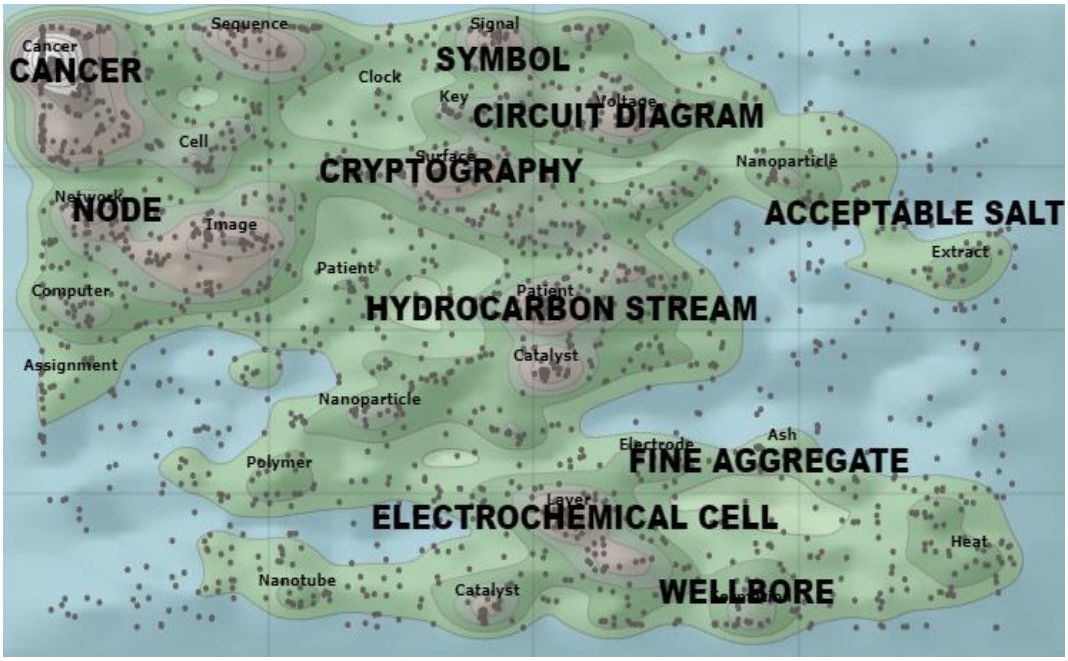


Figure 6. Thematic concept map of inventions by academic institutions in MENA between 2008 and 2021.

The major areas found within the patents dataset of this study include Cancer, SeqID, Node, Symbol, Circuit Diagram, Cryptography, Hydrocarbon stream, Electrochemical Cell, Boiling water, Acceptable Salt, Wellbore, Fine Aggregate, and Exchanger. Some technologies will necessarily overlap, and the delineation of one technical area versus another is therefore only approximative.

Table 1 shows the technologies derived from International Patent Classification (IPC) codes assigned to patents published in the past five years, based on Publication Year. The terms in the Technology column, called 'Smart Themes" supplement the dense IPC definitions with terms derived from actual patents for that technology. These terms are extracted from the DWPI Titles from all patents classified with a specific IPC code. The top key terms are reviewed and represent a clear and concise summary of the technology described by an IPC code. The terms provide fixed descriptions of the technology and do not change based on the patents set. While the technology "*Cancer, Treating, Administering, Disorder, Disease, Inhibitor, Pharmaceutical*" appears twice, these two technologies have different IPC codes, respectively A61K,A61L,C11D and A61P.

Table 1. Top innovations in the past 5 years by academic institutions in MENA by number of patents.

Technology	*Patents*
Cancer, Treating, Administering, Disorder, Disease, Inhibitor,	900
Catalyst, Reactor, Sorbent, Hydrocarbon, Catalytic, Dehydrogenation,	455
Sample, Gas Sensor, Cancer, Cell, Inspection, Antibody, Biological	444
Filter, Membrane, Separation, Gas, Filtration, Carbon Dioxide, Sorbent	337
Surgical, Endoscope, Medical, Patient, Ultrasound, Bone, Tissue	283
Wastewater, Water, Sludge, Desalination, Reverse Osmosis,	262
Computing, Transitory, Touch, Information Processing, User, Virtual,	284
Semiconductor, Layer, Substrate, Oled, Gate, Source Drain, Light	248
Graphene, Carbon Nanotube, Particle, Boron Nitride, Silica, Graphite,	220

Overall, there are 30 different technologies classifications represented in Table 1. The top 3 technologies are found in 24% of the records in the patents dataset of this study. The number of technologies indicates recent innovations and can provide an overview of the current state of the technological market and how it is segmented. These technologies have a direct impact on societal issues related to health (e.g. cancer, treatment, antibody, pharmaceuticals, medical, patient) but also on public environmental issues in the MENA region (water, desalination, purification, filtration). These findings support the hypothesis of this study since Mode 2 research is typically oriented towards practical applications, focusing on solving real-world problems and addressing pressing challenges. It is also the type of research that the industry sector is focused on, often in response to consumer demand.

Co-assignment network and collaboration between Academia and the Industry

This section focuses on the level of co-assignment as a proxy measure of collaboration in patenting activities by the top 18 academic MENA institutions shown in Figure 2 (Saudi Arabian Oil Company and the MIT are excluded). The co-assignment network visualization shown in Figure 7 was created by using VOSviewer at the organization level (van Eck & Waltman, 2009), where a full counting method was used i.e. co-assigned patents are fully assigned to each co-assignee. This network map can also be explored interactively online (https://bit.ly/AcadMENAPatentsMap) and the less visible organizations' names can be seen by zooming in on specific map areas. For readability reasons, the organization name also shows the ISO country code and the colors of the nodes

represent the related countries. The size of the nodes represents the number of patents.

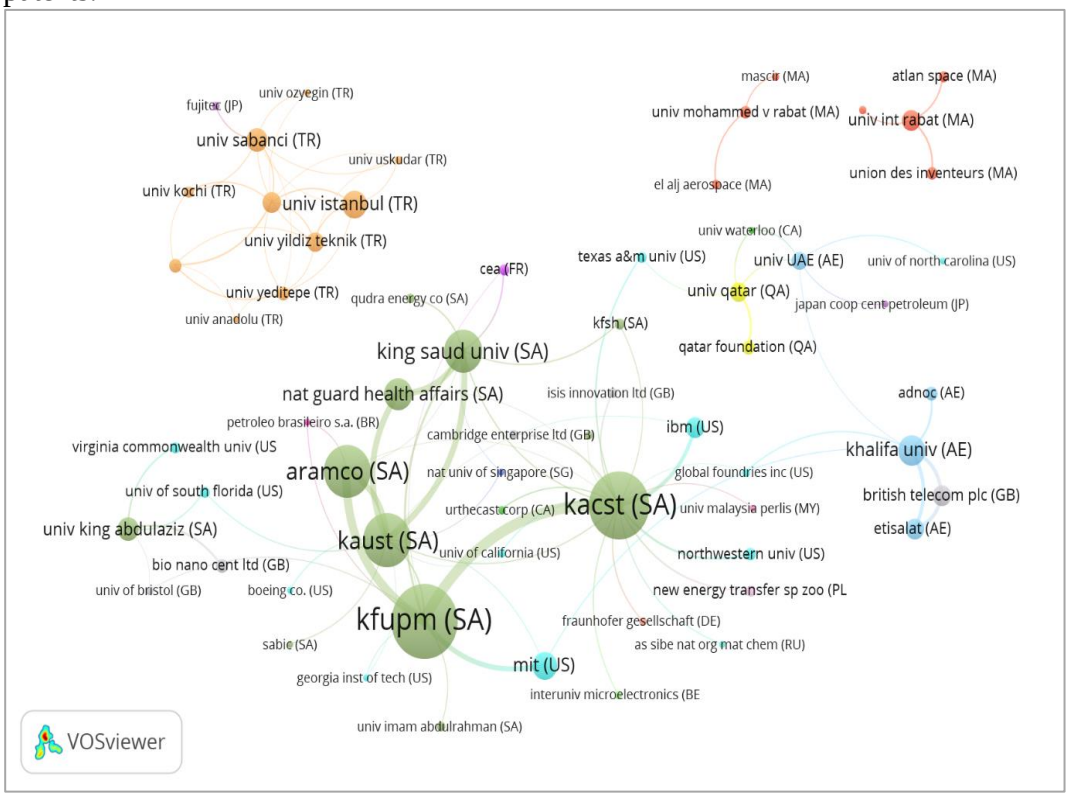


Figure 7. Co-assignment network of the main academic patent assignees in MENA (2008-2021).

These 18 academic institutions contributed to 7,011 inventions (70% of the patents under study). Co-assignments were found in 938 of them (or 13%). In this map, three main areas can be distinguished. On the top left, Turkish academic institutions show a high level of domestic collaboration between academic institutions, and one international co-assignment with a corporation, Fujitec (Japan). On the top right, Moroccan academic institutions show only domestic co-assignments links, including collaborations with local corporations. The third area, shown in the rest of the map, shows the co-assignment links for the institutions in Saudi Arabia (green), UAE (light blue), and Qatar (yellow) which are three countries of the Gulf Cooperation Council (GCC). This area also shows domestic co-assignments but also a much higher level of collaboration with foreign academic and corporate institutions, mainly from the United States (11) and the United Kingdom (5). The co-assignments with domestic corporations include collaboration with Aramco and Sabic in Saudi Arabia, and ADNOC and Etisalat in the UAE. The foreign corporate organizations include Boeing, IBM, British Telecom, Cambridge enterprise, and Petroleo Brasileiro. These findings provide support to the

hypothesis of this study. It is also worth noting that the first two areas of the map are not connected with the third one, which suggests that there is no co-assignment between academic institutions from Morocco and Turkey with their peers in the GCC.

Discussion and conclusion

The original subject of this study was to examine patenting activities of universities in the Middle East and North Africa region. The hypothesis of this study is that there is a positive association between the patenting activities of universities in MENA and the Mode 2 framework of university-industry collaboration as proposed by. To gain a better understanding of the patenting activities in academia within this region, patent data from Derwent Innovation is analyzed to provide key insights on these activities. The findings show that Saudi Arabia lead the MENA region in terms of patent filings with 48% of the patents filled by the academic sector in the region, research institutions in Turkey (28%) and Morocco (11%) follow. The most active academic institutions in patenting activity are located in Saudi Arabia, Turkey, Morocco, UAE and Qatar. The number of patents grew by 4,604% between 2008 and 2021 for MENA academic institutions compared with a growth of 2,329% for academic institutions worldwide. The patenting activity by the academic sector grew faster in MENA compared to the World, but the region started from a relatively low base in 2008. Patent protection continues to be most often sought in the United States, and the Patent Cooperation Treaty (PCT) application route is also commonly used by academic institutions in MENA. The largest technical fields of the patents include the Polymers & Plastics, the Pharmaceuticals and Computing & Control. Some of the underlying technologies have a direct impact on societal health-related issues (e.g. cancer, treatment, antibody, pharmaceuticals, medical, patient) but also on public environmental issues (water, desalination, purification, filtration). These main academic assignees show a certain level of domestic and international collaboration with other academic institutions but also corporations. More specifically, academic institutions in Saudi Arabia, the UAE and Qatar show linkages with the industry sector which might suggest a certain potential in terms of commercialization of research done by the academic sector on practical applications and solutions to real-world problems.

This study also contributes to a more inclusive assessment of research produced in MENA by academic institutions as it includes economic and societal dimensions of research activities. Indeed, it covers a different type of research activities beyond journal publications and practices such as patenting activities and collaboration with the industry. This study provides also insights about valuable contributions that researchers in MENA make to science for the benefit of society. The growth of patenting activities in MENA may seem impressive on a standalone basis, but

when compared to the level of innovation worldwide, the region still lags behind the rest of the world. Corporates are more likely to invest in innovation when there is more patent protection (Allred & Park, 2007) and might collaborate with the Academic sector more frequently. The private sector in MENA might be encouraged to boost its patenting activity thanks to relevant national legislations that are consistent with global best practices. Due to its indirect relation to technical innovation, current government policies and funding processes to support academic research alone in MENA may not be the best mechanisms to develop further the patenting activities by research institutions. The ability to commercialize a product, typically accomplished by corporations, and collaborations with the industry are likely to be the major driving forces behind an increase in patenting in the region by the academic sector.

Another theoretical framework that could be incorporated into a future study is the Triple Helix concept which proposes a collaborative and dynamic relationship between the government, academia, and industry sectors (Etzkowitz & Leydesdorff, 1995; Leydesdorff & Meyer, 2007). According to the Triple Helix model, all three sectors play important, complex and interrelated roles in the innovation process, with knowledge, resources, and benefits flowing in multiple directions between the different sectors. The Triple Helix model acknowledges the strengths and perspectives of each sector. Academia is typically responsible for the creation of new knowledge; the government sector shapes the broader policy and regulatory landscape and the industry sector is focused on the practical application of research and innovation. To better understand the relationship between government policies and technology development in MENA, future research could focus on various aspects such as national regulatory frameworks, investment incentives, and intellectual property rights. More specifically, future studies may explore the effectiveness of these policies and identify potential trade-offs or synergies between different objectives such as economic growth, social welfare, and environmental sustainability. Another research opportunity consists of examining how policy design and implementation vary across different political regimes and institutional contexts within the MENA region, and whether there are any lessons that can be drawn from successful cases in other regions or countries.

Acknowledgments

I am particularly grateful to Ludo Waltman for his useful comments and suggestions on an earlier version of this manuscript. I would like to thank the two anonymous reviewers for their comments and suggestions.

Competing Interests

The author is an employee of Clarivate Analytics, the provider of Derwent Innovation.

References

Allred, B. B., & Park, W. G. (2007). The influence of patent protection on firm innovation investment in manufacturing industries. *Journal of International Management, 13*(2), 91-109. https://doi.org/https://doi.org/10.1016/j.intman.2007.02.001

Altbach, P. G. (2009). Peripheries and centers: Research universities in developing countries. *Asia Pacific Education Review, 10*, 15-27.

Basberg, B. L. (1987). Patents and the measurement of technological change: a survey of the literature. *Research Policy, 16*(2-4), 131-141. https://doi.org/10.1016/0048-7333(87)90027-8

Clarivate. (2022). *ThemeScape*. https://www.derwentinnovation.com/tip-innovation/support/help/themescape.htm

Clark, B. R. (1998). *Creating entrepreneurial universities: organizational pathways of transformation. Issues in Higher Education*. ERIC.

Directorate-General for Research and Innovation. (2022). Reforming research assessment: The Agreement is now final. https://research-and-innovation.ec.europa.eu/news/all-research-and-innovation-news/reforming-research-assessment-agreement-now-final-2022-07-20_en

Dornbusch, F., Schmoch, U., Schulze, N., & Bethke, N. (2013). Identification of university-based patents: A new large-scale approach. *Research Evaluation, 22*(1), 52-63. https://doi.org/10.1093/reseval/rvs033

Etzkowitz, H., & Leydesdorff, L. (1995). The Triple Helix--University-industry-government relations: A laboratory for knowledge based economic development. *EASST review, 14*(1), 14-19.

Hanafi, S., & Arvanitis, R. (2015). *Knowledge production in the Arab World: the impossible promise*. Routledge.

Landini, F., Malerba, F., & Mavilia, R. (2015). The structure and dynamics of networks of scientific collaborations in Northern Africa. *Scientometrics, 105*(3), 1787-1807. https://doi.org/10.1007/s11192-015-1635-1

Leydesdorff, L., & Meyer, M. (2007). The scientometrics of a Triple Helix of university-industry-government relations (Introduction to the topical issue). *Scientometrics, 70*(2), 207-222.

Limoges, C., Scott, P., Schwartzman, S., Nowotny, H., & Gibbons, M. (1994). The new production of knowledge: The dynamics of science and research in contemporary societies. *The New Production of Knowledge*, 1-192.

Lissoni, F. (2013). Academic Patenting in Europe: A Reassessment of Evidence and Research Practices. *Industry and Innovation, 20*(5), 379-384. https://doi.org/10.1080/13662716.2013.824190

Mejlgaard, N., & Ryan, T. K. (2017). Patterns of third mission engagement among scientists and engineers. *Research Evaluation, 26*(4), 326-336. https://doi.org/10.1093/reseval/rvx032

Noruzi, A., & Abdekhoda, M. (2012). Mapping Iranian patents based on International Patent Classification (IPC), from 1976 to 2011. *Scientometrics, 93*(3), 847-856. https://doi.org/10.1007/s11192-012-0743-4

Pavitt, K., & Walker, W. (1976). Government policies towards industrial innovation: a review. *Research Policy, 5*(1), 11-97.

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Brostrom, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., Krabel, S., Kitson, M., Llerena, P., Lissoni, F., Salter, A., & Sobrero, M. (2013). Academic engagement and commercialisation: A

review of the literature on university-industry relations. *Research Policy*, *42*(2), 423-442. https://doi.org/10.1016/j.respol.2012.09.007

Tödtling, F., & Trippl, M. (2005). One size fits all?: Towards a differentiated regional innovation policy approach. *Research Policy*, *34*(8), 1203-1219. https://doi.org/10.1016/j.respol.2005.01.018

Uzun, A. (2001). Technological innovation activities in Turkey: the case of manufacturing industry, 1995-1997. *Technovation*, *21*(3), 189-196. https://doi.org/10.1016/s0166-4972(00)00033-x

van Eck, N. J., & Waltman, L. (2009). VOSviewer: A Computer Program for Bibliometric Mapping.*Proceedings of the International Conference on Scientometrics and Informetrics* Proceedings of Issi 2009 - 12th International Conference of the International Society for Scientometrics and Informetrics, Vol 2, Leuven.

van Zeebroeck, N., de la Potterie, B. V. P., & Guellec, D. (2008). Patents and academic research: a state of the art. *Journal of Intellectual Capital*, *9*(2), 246-+. https://doi.org/10.1108/14691930810870328

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. https://doi.org/10.13140/RG.2.1.4929.1363

World Bank. (2019). *Middle East and North Africa*. World Bank. https://www.worldbank.org/en/region/mena

https://doi.org/10.51408/issi2025_039

Exploring Multi-Energy Convergence Through Knowledge Graphs and Patent Bibliometrics

Shuying Li[1], Xian Zhang[2], Wudan Ma[3], Xiaoyu Wang[4], Chunjiang Liu[5], Haiyun Xu[6], Edwin Garces[7]

[1]*lisy@clas.ac.cn*, [2] *zhangx@clas.ac.cn*, [4] *wangxy@clas.ac.cn*, [5] *liucj@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610029 (China)

[3]*mawudan1997@163.com*
Zhejiang Normal University Library, Jinhua 321000 (China)

[6]*xuhaiyunnemo@gmail.com*
Shandong University of Technology, Zibo City, 255000 (China)

[7]*edwingus@gmail.com*
Portland State University, 1825 SW Broadway, Portland, OR 97201 (United States)

Abstract

Today, low-carbon, clean energies - renewables, hydrogen and nuclear - have begun to replace fossil fuels. This transition has been accompanied by an integration of new energy technologies in terms of shared use of energy, integration of multiple energy systems, and conversion between energy sources. Countries are actively building a low-carbon energy system with multi-energy integration to achieve the dual-carbon goal. This paper proposes to construct a multi-energy patent knowledge map and establish a domain knowledge organizing system based on fusing multiple technology classifications by integrating different patent technology classifying systems. Top-down and bottom-up approaches are adopted to build a conceptual model, and empirical research and validation are conducted in the field of low-carbon energy technology as an example to systematically analyze the development trend of low-carbon energy, convergence signals, and the potential of multi-energy convergence. The results are expected to provide insights into the development and practical application of multi-energy technologies and provide a basis for formulating relevant policies and research directions.

Introduction

Reducing carbon emissions to combat climate change is becoming a global consensus, and "dual carbon" (The goals for peak CO2 emissions and carbon neutrality) is an important strategic goal for most countries around the world in the next half century. Improving the energy supply structure is the linchpin and key to realizing the "dual carbon" path. In order to achieve this objective, it is imperative to gradually and steadily transition away from a coal-based energy structure towards a more robust and diverse energy portfolio. This transition requires the vigorous development of both renewable energy sources and safe and advanced nuclear energy. Additionally, it is essential to recognize the complementarity and large-scale potential of non-fossil energy sources, fostering a multifaceted approach to energy production and consumption. Presently, an array of low-carbon and clean energy sources is emerging. Such resources include photovoltaic, solar thermal, wind, nuclear, hydrogen, biomass, ocean energy, and geothermal energy. However, renewable energy is faced with significant challenges, including low energy density,

high volatility, intermittent availability, and inherent randomness. Consequently, the implementation of renewable energy on a large scale necessitates a systematic integration of diverse energy sources within the overall energy system. For instance, wind and light resources can serve as the primary sources of power generation and energy supply, whereas nuclear power, hydropower, and analogous comprehensive and complementary non-fossil energy sources can be utilized as a "stable power source," with a modicum of thermal power functioning as an emergency power source or a regulating power source. The development of a new type of power system management and operational framework will be enabled by the integration of renewable energy power prediction technology, advanced power system stabilization and control technology, and innovative power system flexible interaction technology. Beyond electrochemical energy storage, mechanical energy storage, electromagnetic energy storage, and hydrogen energy, a broad range of energy storage methods is considered. Consequently, the focal point of establishing a multi-energy complementary integrated energy system is the mastery and realization of the core technology of multiple energy coupling and complementary (Li et al., 2022).

Technological convergence is the process of combining existing technologies into hybrid technologies (Curran, Bröring and Leker, 2010). This integration is not just about adding technology but innovating in unprecedented ways to create new markets. The convergence of technologies for different new energy sources encompasses the joint utilization of energy resources, the integration of multiple energy systems, and the interconversion of energy sources. For instance, the technology of using nuclear energy to produce hydrogen energy has emerged as a promising avenue for the future, offering a carbon-free approach to hydrogen production. Significantly, prominent developed nations such as the United States and the United Kingdom have unveiled comprehensive research and development plans with the objective of fostering the advancement and integration of these technologies, such as US's Nuclear Hydrogen R&D Plan (DOE, 2022) and the report Unlocking the UK's Nuclear Hydrogen Economy to Support Net Zero (National Nuclear Laboratory, 2021). Hydrogen applications aim to explore flexible and efficient multi-energy integration solutions while enhancing the performance of existing fuel cell systems (Yue et al., 2021). As hydrogen energy continues to be developed, its applications are expected to evolve from single solutions to composite systems. Examples include pathways from renewable power generation to hydrogen, methanol, and chemical feedstocks; and systems from electricity to hydrogen and power for use in exploring multi-energy integration based on hydrogen energy (Fu et al., 2020). In this paradigm, hydrogen emerges as a pivotal energy carrier, facilitating flexible complementarity among diverse energy sources and promoting decarbonization in multiple sectors, including power, transportation, chemicals, and steel, through conversion to electricity, heat, gas, or as raw materials (Li, He and Farjam, 2023).

Achieving carbon neutrality depends on the widespread adoption of renewable energy and new energy technologies. However, large-scale deployment of renewable energy is challenging. In particular, the synergistic and interactive use of different energy resources is crucial. To overcome these challenges, renewable energy sources

such as wind and solar must be integrated with stable energy sources such as nuclear, hydro, and other non-fossil fuel sources. Thermal power can be used as an emergency backup. It is necessary to develop a new framework for managing and operating power systems. To do this, it is important to understand how different new energy sources can be integrated. Based on the multi-energy patent knowledge graph, we analyze the trends, convergence signals, potential and evolution paths of multi-energy integration using patentometrics. The research questions are as follows: What are the developments in the integration of different types of renewable energy sources? What is the potential for integrating multiple renewable energies? How does it work to integrate multiple renewables? What is the direction of low-carbon energy technology integration and technology evolution path?

This paper explores the domain knowledge discovery of technological convergence, proposing a method and process for doing so based on a convergence perspective. The investigation and analysis of existing patent technology classification systems and industrial classification systems worldwide is initiated to establish a foundation for the subsequent analysis. The design of an automatic mapping model of multiple classifications employs the integration of conceptual-level and data-level knowledge, aiming to merge disparate patent technology classification systems and construct a domain knowledge organization system. The proposed methodology integrates a top-down (knowledge conceptualization) and bottom-up (knowledge refinement) approach, facilitating the identification of domain knowledge. The top-down approach of knowledge concept refinement is integrated with the bottom-up approach of entity category summarization to construct the conceptual model of domain knowledge mapping. The constructed method is then applied to investigate technological innovation opportunities and evolution paths. An experimental study is conducted in the field of low-carbon energy technology to verify the feasibility and validity of the constructed methodology and process.

This paper focuses on two main aspects of technological integration and development trends in major low-carbon energy technologies. Firstly, it analyses the technological integration trend of various low-carbon energy technologies based on a multi-energy patent knowledge map. Secondly, it clarifies the main technological direction and evolutionary path of multi-energy integration. The primary objective of this study is to provide a comprehensive basis for the formulation of relevant policies and research directives.

Literature Review

Renewable Energy and Patent Classification

Major national intellectual property offices and organizations in the world have established patent classification systems in the field of renewable energy, covering the concept of "multiple energy sources" and the classification system (**Error! Reference source not found.**), including seven types of renewable energy sources, such as solar, wind, nuclear, hydrogen, biomass, ocean, geothermal and other renewable energy sources. The patent classification system related to green transition technology is a low and zero carbon energy-related technology classification or

patent search formula formed through the discussion of experts in the field, which facilitates the wider use of it to conduct patent information analysis. In this context, WIPO has created a patent classification index for climate change mitigation technologies that are consistent with the existing International Patent Classification (IPC) system. China and Japan, which are relatively late in adopting it, are formulating it from an energy supply and utilization from industry and electric power generation perspective. The newly established Y02E (low-carbon energy generation, transmission and distribution technologies) in the Joint Patent Classification System for European-American cooperation. The diversification of classification systems has two notable effects. On the one hand, it provides richer paths for accessing information. On the other hand, it significantly increases the uncertainty factor. In cases where multiple knowledge sources correspond to the same technical feature, each source may adopt a different technical classification and attribute framework. This often leads to fragmentation of knowledge organization. As a result, issues such as knowledge redundancy, semantic ambiguity, and inconsistent quality may arise. These problems exacerbate the uncertainty in the knowledge acquisition process. They also challenge the reliability and confidence level of the knowledge. In this context, the effective integration and fusion of multi-source knowledge have become essential strategies for enhancing the accuracy of knowledge discovery.

The central objective of this section of the study is to employ conceptual-level knowledge fusion techniques, with the aim of integrating the same knowledge source—which utilizes different classification and attribute systems—into a unified global framework. This process focuses on solving key issues such as conflict detection, entity disambiguation, entity alignment, and collaborative reasoning that arise when different classification systems point to the same knowledge content. It also lays a solid foundation for the seamless integration of multiple technology classification systems.

The paper systematically summarizes and reorganizes seven categories of low-carbon energy, including major technology categories, industry divisions, and domain-specific classifications, which are then further linked to the corresponding entries in the International Patent Classification (IPC) and the Cooperative Patent Classification (CPC) systems. This process of summarization and reorganization serves to enhance the coherence and consistency of knowledge representation. Moreover, it provides a clearer and more comprehensive perspective for the subsequent analysis of technology integration and innovation.

Table 1. Patent Classification of Multiple Energy Sources.

Organization	*Patent Classification*	*Different Definition*
CNIPA (CNIPA, 2023)	Patent Classification System for Green and Low Carbon Technologies	Fossil Energy Carbon Reduction; Energy Conservation and Recycling; Clean Energy; Energy Storage; CCUS
WIPO (WIPO, 2010)	WIPO IPC Green Inventory	Nuclear power generation, alternative energy (biofuels, fuel cells, hydrogen, wind, solar, geothermal, waste heat, etc.)
USPTO EPO (USPTO; EPO, 2010)	CPC classification	Y02E (low-carbon technologies related to energy production, transmission and distribution) Y02E10/1 (geothermal), Y02E10/2 (hydro), Y02E10/3 (ocean), Y02E10/4 (solar thermal), Y02E10/5 (photovoltaic), Y02E10/7 (wind), Y02E50/1 (biofuels), Y02E50/3 (waste fuels), Y02E30/1 (spent biofuels), Y02E30/1 (nuclear fusion), Y02E30/3 and Y02E30/4 (nuclear fission)
USPTO (USPTO, 2009)	EST Concordance	Alternative energy: biomass, fuel cells, geothermal energy, hydroelectric energy, solar energy, wind energy
JPO (JPO, 2022)	Green Transformation Technologies Inventory	Energy Supply (gxA): Photovoltaic Power Generation, Solar Thermal Power Generation, Wind Power Generation, Geothermal Power Generation, Hydropower, Ocean Energy Power Generation, Biomass, Nuclear Power Generation, Fuel Cells, Hydrogen Technology, Ammonia Technology

Low-carbon technologies and Patent Analysis

Patent bibliometrics is an important method for studying the innovative output of low-carbon technologies. Analyzing low-carbon energy patent data can provide insights into the development trends and trajectories of low-carbon technologies. Oltra and Saint Jean (2009) argues that patents are a useful means of measuring green energy technologies. They can analyze invention activities in specific technological fields, the international dissemination of technology, the research and technological capabilities of enterprises, and the sources of knowledge of innovative institutions, as well as technological spillovers. Albino et al. (2014) analyzed the development

and impact of low-carbon energy technologies, examining nuclear power production, alternative energy production, and energy conservation patents in The IPC Green Inventory, and found that the United States is the main source of innovative low-carbon energy technologies, while Japan leads in solar energy and low-energy lighting. Although China, Russia, and other countries are increasingly using low-carbon energy technologies, the level of technological innovation in this area remains low. Leu, Wu and Lin (2012) analyzed the status of technology development in the field of biofuel and biohydrogen energy on the basis of patentometrics, and found that the U.S. is leading the development of biofuel-related energy, and the high number of cited patents suggests that biofuel production technology must give priority to low energy demand. Liu et al. (2011) classified patents related to photovoltaic technology based on keyword co-occurrence and analyzed the growth trajectory of five groups of photovoltaic technologies. Chen, Chen and Lee, (2011) conducted a bibliometric and patent analysis to study the technological evolution and patent strategy of hydrogen energy and fuel cells. Subtil Lacerda (2019) examines the influence of scientific knowledge on the evolution of wind turbine technology trajectories through bibliometric analysis and finds a strong correlation between the development of scientific knowledge and the technological trajectories of wind turbines. Similarly, Hötte, Pichler and Lafond (2021) analyzed the relationship between low-carbon energy technologies and scientific knowledge. By analyzing a corpus of patents covering six renewable energy technologies from 1970 to 2019, Jiang et al.(2022) sheds light on the life cycle of these technologies, the technological landscape, the potential markets, and the competitive landscape in key countries/regions involved. The current study is mainly a descriptive analysis of LCE, which is limited by data availability and data processing capabilities. Second, the static patent classification system on which LCE is based is not perfect. It does not analyze trends in multi-energy convergence. The boundaries between fields are not clear and may evolve with dynamic cross-field convergence.

Second, technology convergence research based on patent information has become the main method and hot direction of technology convergence research. In addition, there are related studies that use data from papers, standards, and Wikipedia. For measuring technology convergence, Herfindahl Index, patent cross-impact analysis, social network analysis, and time window analysis are applied(Jeon and Suh, 2019; Lee, Kogler and Lee, 2019). In predicting technology convergence trends, methods such as link prediction based on technology convergence networks(Park and Yoon, 2018), neural network method based on technology convergence matrix (Kim and Lee, 2017), and time series prediction method based on time series of technology convergence relationships(Lee, Park and Kang, 2018) have been applied. Xue and Shao (2024) identifies technological evolution paths in the field of hydrogen energy using patent text mining. The analysis shows a good convergence in the evolution of hydrogen energy technologies, focusing mainly on hydrogen storage materials, hydrogen fuel cell vehicles, and green hydrogen production. Existing research shows that technology convergence positively affects technology value and innovation activity. Most empirical studies, however, are highly generalizable across domains. Few studies analyze the dynamics of technology convergence for multiple domains

and across domains. At present, there are fewer studies on technology fusion analysis for patent data that target multi-domain and domain-wide technology fusion dynamics. This study attempts to fill this research gap by constructing a multi-energy fusion patent knowledge map.

The Knowledge Graph (KG) and Knowledge Discovery

The Knowledge Graph is a structured Semantic Web knowledge base that describes concepts and how they relate to each other in a visual way by mapping abstract data and knowledge to graphical elements(Dessì *et al.*, 2021). It complements human-computer interaction by helping users effectively perceive and analyze data and knowledge while exploring connections to extend existing knowledge(Xiao, Li and Thürer, 2023). It can mine and analyze knowledge and its interrelationships and are important tools for paying attention to the frontiers of science and technology and knowledge management. The existing studies mainly focus on the concept, development history, structure, application and so on aspects of knowledge maps(Nguyen and Chowdhury, 2013; Balaid *et al.*, 2016). Based on co-word analysis, social network analysis and strategy analysis, Pino-Díaz *et al.*, (2012)proposed the method of constructing techno-scientific network strategic knowledge map, which can visualize strategic knowledge, keywords, subnetwork proximity and other contents. Su and Lee, (2010)proposed a three-dimensional network and a two-dimensional map based on the co-occurrence of keywords, which can describe the forward-looking knowledge structure of the latest technology in a quantitative and visual way. The concept of a knowledge graph remains undefined, and research in this area is still in its early stages. Most researchers are now building knowledge graphs as navigational aids (network analysis, visualization, or text mining, etc.), which play an important role in organizing knowledge acquisition, connecting experts, discovering knowledge, and facilitating mobility (Lee and Fink, 2013). Key challenges include domain knowledge organization, dynamic/tacit knowledge representation/extraction, and cross-domain knowledge mapping (Suresh and Egbu, 2004). Zhou *et al.* (2024) maps knowledge on hydrogen fuel cell technology on the basis of bibliometrics and IPC co-classification analysis.

Karlapalem (2021) believes that Knowledge Discovery in Database is an important process for identifying valid, novel, potentially useful and ultimately understandable patterns in data, which refers to the extraction of implicit, unknown and potentially useful information from data (Fayyad, 2001), and the term refers to research results, technologies and tools that extract useful information from a large amount of data(Agrawal and Shafer, 1996). The extracted information includes concepts, relationships between concepts, classifications, decision rules and other information(Vickery, 1997). Knowledge discovery emphasizes that knowledge is the product of data-driven discovery process, a common point of different research fields, focusing on data analysis and knowledge extraction from different perspectives, such as database, statistics, mathematics, logic or artificial intelligence (Mariscal, Marbán and Fernández, 2010). Due to the complexity of knowledge and the Fusibility of technologies, it is very important to adopt appropriate methods and perspectives for knowledge discovery and analysis. In recent years, various

knowledge discovery methods have been rapidly developed and widely applied in various industries, such as cancer diagnosis, biological classification of river water quality, population analysis, quality control, disaster risk assessment, global climate change modeling, time series pattern analysis, clinical medicine (Sebastian and Then, 2011; Anguera et al., 2016), topology optimization (Yamasaki, Yaji and Fujita, 2019), etc. Roscher et al.(2020) analyzed the application of explainable machine learning in natural science, holding that its main goal is to obtain new scientific insights and discoveries from observation or simulation data, and that the prerequisite for obtaining scientific results is domain knowledge, and defined the concepts of transparency, interpretability, and explainability.

Existing knowledge discovery techniques, research methods, perspectives, and outcomes are increasingly exhibiting a trend toward diversification. To accurately describe and reveal the knowledge structure and evolution characteristics, and to avoid discovering local and one-sided knowledge, it is necessary to integrate heterogeneous data from multiple sources. Furthermore, the knowledge organization system must be improved to maximize the discovery of the domain knowledge structure and the dynamic evolution characteristics from the perspective of data and technology convergence.

Methodology

The process consists of three main steps. First, we develop a system to organize domain knowledge using a technology classification framework derived from various sources. Next, we construct a knowledge graph. Finally, we leverage a multi-energy knowledge graph to conduct empirical research on technology convergence.

Domain knowledge under a multi-source technology classification system

This study employs conceptual-level knowledge convergence, a process aimed at integrating knowledge sources from various classification and attribute systems into a unified global framework. The theoretical underpinnings of knowledge classification and fusion of multi-feature representations (see Fig. 1) serve as the foundation. Initially, the knowledge system of the domain is extracted. Then, through the direct merging of the extracted data, the representation of "concept, attribute and attribute value" is formed (e.g., wind energy, IPC/CPC).

Subsequently, the entity references are categorized based on the established classification and fusion rules. The specific principles that have been adopted are as follows:

First, classification principles: (1) concept mutual exclusion constraint, i.e., the more intersected, the more compatible the concepts; (2) hierarchical concept constraint, i.e., an entity does not belong to a certain concept, and it does not belong to any sub-concepts.

Second, fusion principles: (1) concept fusion, which refers to synonyms or similar concepts; (2) attribute alignment, i.e., the degree of overlap of entity-attribute values corresponding to attributes; and (3) attribute value alignment, i.e., deletion of duplicates and elimination of erroneous knowledge.

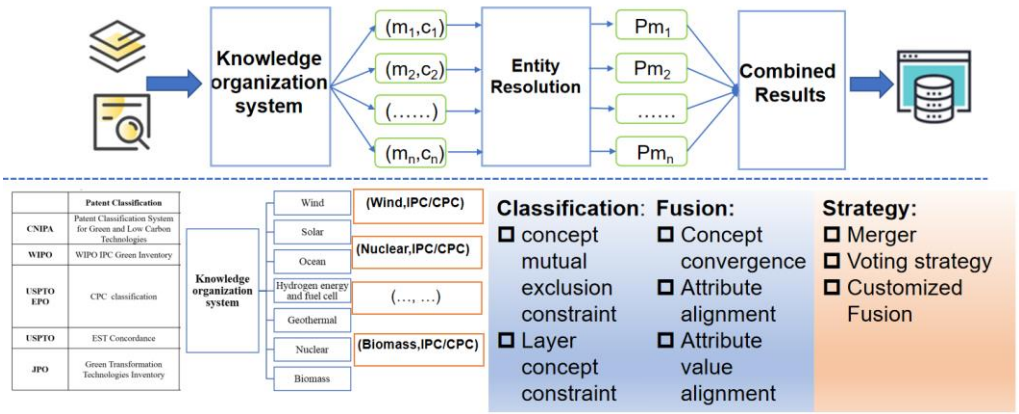


Figure 1. Theoretical Foundations of Knowledge Classification and Fusion Based on Multi-Feature Representation.

The theoretical foundation outline above serves as the basis for the development of the automatic mapping model of classes for multiple classifications (Fig.2). This model specifically incorporates two levels of patent classification feature fusion methods.

(1) Concept layer convergence

The first layer involves the convergence of text-based and structure-based approaches. The text-based method entails matching of the ontologies through the textual description information, the extraction of the descriptions from two ontologies, and the similarity between them. The structure-based method utilizes structural information between the ontology concepts when the textual information is inadequate for determining the matching relationship between two ontologies.

Initially, the text in "concept, attribute and attribute value" is extracted, including the text of technical categories, explanations, and IPC classification descriptions. Subsequently, the extracted text information is used to map into various vectors that can be corresponded to, in the form of, e.g. $\overline{\text{wind}} = \left(w_{i,1}, w_{i,2}, w_{i,3}, \cdots, w_{i,n}\right)$, which is the set of vectors. Thirdly, the semantic similarity between the vectors is calculated by using the cosine similarity, the Euclidean distance, and other metrics. The calculation of semantic similarity between the vectors is performed using cosine similarity, Euclidean distance, and other metrics.

(2) Data layer convergence

An instance-based approach is adopted in this context. The instances of ontology concepts are utilized as the basis for similarity measurement when calculating ontology similarity. The number of identical instances of two ontologies is compared to calculate the similarity between ontologies. The greater the similarity, the closely the two ontologies align. This method is highly reliable.

The specific operational procedure is outlined as follows: Initially, the IPC classification number and attribute value corresponding to the ontology are extracted. Subsequently, under specific conditions, the probability model is employed to ascertain the matching relationship between the entities in question, i.e.,

the IPC classification number and other entities (single patents) with an IPC relationship.

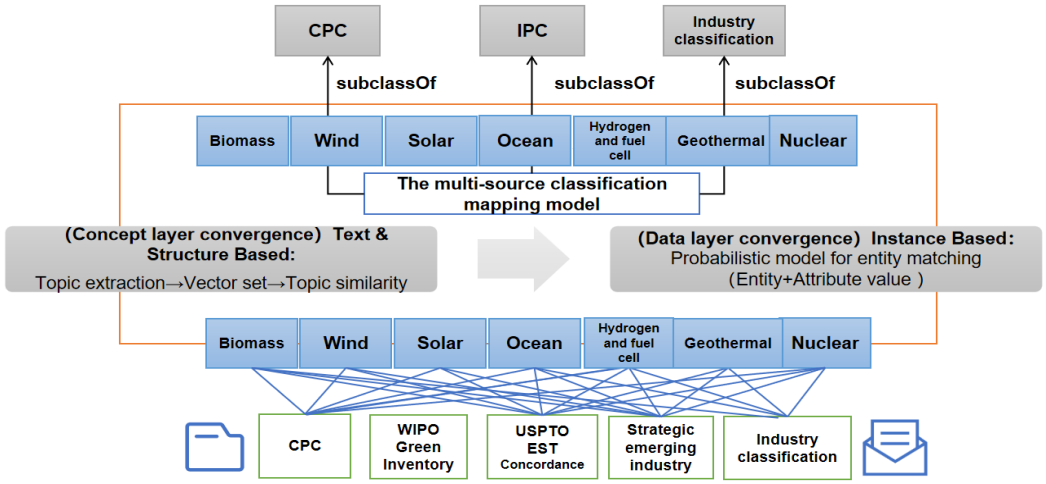


Figure 2. The multi-energy classification mapping model.

Multi-energy Knowledge Graph

The conceptual model of Multi-energy Knowledge Graph is constructed by combining top-down and bottom-up approaches (Fig.3). Firstly, top-down approach utilizes the knowledge organization system to gradually refine the concepts from the top level down to form a tree-structured mapping model. Secondly, bottom-up approach uses the patent data, which has been summarized by the related entity categories, to form a broad category scope layer from multiple fields of patents upward, thereby forming a general patent knowledge graph. The final step involves the combination of the two approaches to form a generic patent knowledge mapping by means of attribute extraction, attribute alignment, relationship construction, concept hierarchy construction and entity classification, etc., to obtain data and realize the construction of domain-specific knowledge graph.

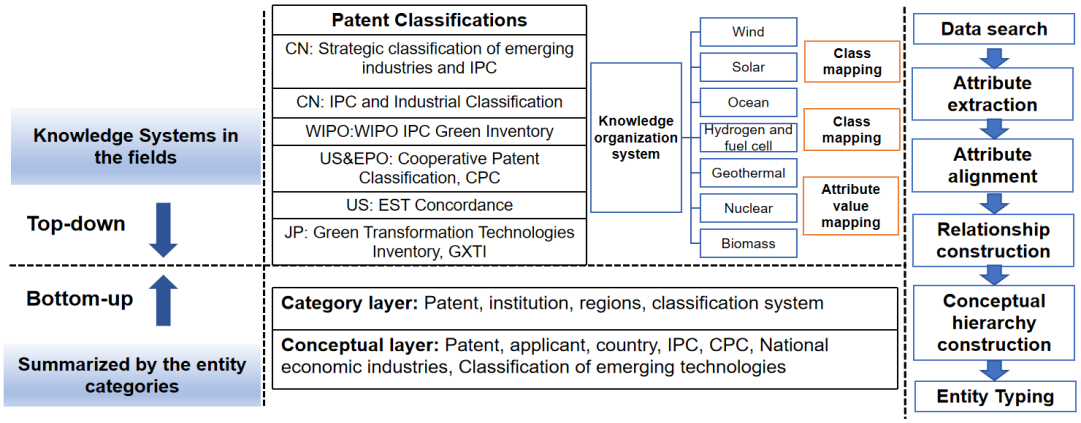


Figure 3. Method for constructing multi-energy knowledge graph.

Data search strategy

The data source of this study is the emission peak and carbon neutrality patent information platform (www.cpnp.ac.cn) built by the Chinese Academy of Sciences based on the web, which is a one-stop and patent big data information service platform, and contains a large amount of emission peak and carbon neutrality patent information worldwide. This platform has strong professional relevance and supports the comprehensive collection of patent data relevant to various energy sources, which is highly consistent with the research topic. In this paper, we searched for priority patents related to low-carbon zero-carbon energy, energy storage and multi-energy integration technologies to ensure the timeliness and novelty of the data. The search results involved a total of 7 technology branches, and obtained more than 1.5 million pieces of relevant patent data (Table 2). The search was conducted in April 2022.

Table 1. Patent search strategy and data proportion.

Fields	*Secondary Fields*	*The Number of Patents*	*Proportion (%)*
Low-carbon and zero-carbon energy (1,041,553)	Nuclear power and non-electric use of nuclear energy	131,903	12.6%
	Renewable energy (Solar, wind energy, biomass energy, geothermal, ocean energy)	716,720	68.5%
	Hydrogen energy and fuel cell	198,266	18.9%
Energy storage and multi-energy integration	Heat/cold storage	55,065	10.2%
	Physical power storage	44,051	8.2%
	Chemical power storage	416,783	77.5%
	New power systems based on renewable energy	21,605	4.0%

Technical comparison indicators

We combines the characteristics of different technical fields and introduces the following two technical comparison indicators:

(1) Technology comparative advantage

Low-carbon clean energy technologies can be categorized into the following: renewable energy, hydrogen and fuel cells, nuclear power, and non-electric utilization of nuclear power. Energy storage and multi-energy integration technologies can be categorized into the following: thermal energy storage, new power systems based on renewable energy, chemical power storage, and physical power storage. The patent technology dominance of country j in the ith technology field (second level) can be calculated by formula (1-1) using the internationally recognized multi-disciplinary measurement index "technology comparative advantage" (RTA).

$$RTA = \frac{P_{ij} / \Sigma_i P_{ij}}{\Sigma_j P_{ij} / \Sigma_i P_{ij}} \quad (1\text{-}1)$$

In equation, P_{ij} denotes the number of patents of the jth country in the ith technology field.

(2) Technological relevance

The integration of wind energy with other energy technologies has become increasingly prominent. The coefficient of technological relevance is employed to assess the technological relevance of the seven energy sources. Compared with indicators such as Jaccard Index or Salton Cosine, which can only capture the differences between technologies, the correlation coefficient can capture the distance between two technologies and is more conducive to evaluating the closeness of the relationship between technologies. A larger value indicates a closer relationship between the two technologies. The calculation method is delineated in equation (1-2):

$$S_{ij} = \frac{\sum_{n=1}^{k} C_{in} C_{jn}}{\sqrt{\sum_{n=1}^{k} C_{i\,n^2}} \sqrt{\sum_{n=1}^{k} C_{j\,n^2}}} \quad (1\text{-}2)$$

In equation (1-2), S_{ij} denotes the correlation coefficient between technologies i and j. If S_{ij} is equivalent to 1 on the diagonal of the correlation matrix, it signifies that the co-occurrence distribution of technologies i and j in patents is entirely consistent, indicating a complete integration of each technology with itself. Conversely, if S_{ij} is 0, it indicates that the distribution of patents for technologies i and j is entirely disjoint. K represents the number of core technologies, which is to say the width of the integration of the technologies is represented by Cjn, which denotes the number of instances in which technologies j and n are present together in a single patent.

Empirical Study of Convergence application of Multi-energy Knowledge Graph

The domain of low and zero-carbon energy technologies has been selected as a subject of in-depth experimental research. This decision stems from two primary considerations. Firstly, green and low-carbon technologies are garnering increased global attention, prompting prominent scientific and technological powerhouses, as well as regional organizations, to dedicate considerable resources to the promotion of research and development in the field of green technologies. Notably, the development of patent-technology classification systems, a crucial component in the

patent information search strategy, is experiencing robust growth. This system is instrumental in facilitating a comprehensive and systematic understanding of the patent landscape, thereby providing a solid knowledge framework and conceptual foundation for the execution of this research method.

Secondly, international science and technology and industrial communities have urgent strategic needs for low-carbon energy and multi-energy fusion technologies. An in-depth investigation of this technological innovation trend in the field can track the development of innovation and application among low-carbon energy supply technologies. It can also quickly capture the characteristics of technology convergence and its path evolution trends. This is imperative for the development of a low-carbon energy system integrating various energy sources and for providing a scientific foundation for related innovation decisions.

Knowledge Organization for Multi-energy Classification

In view of the major strategy of low-carbon energy and the demand for multi-energy integration, the concept of "multi-energy" and its classification system have been investigated. A comparative analysis demonstrates that the World Intellectual Property Organization (WIPO) and the United States Patent and Trademark Office (USPTO) have pioneered the Green Patent Technology Classification System, with the classification primarily devised from an alternative energy perspective. In contrast, the Joint Patent Classification System of Europe and the United States has recently augmented its classification with the addition of category Y02E (low-carbon technologies associated with energy generation, transmission, and distribution). Meanwhile, China and Japan have demonstrated a bias towards the perspectives of energy supply and utilization in the new energy industry and the electric power production industry (Guo R, et al., 2020).

The present study, which is based on the patent technology classification systems mentioned above, draws upon the technical characteristics of the respective fields. It identifies and explores the knowledge concept space of seven types of low-carbon clean energy, such as solar, wind, nuclear, hydrogen, biomass, ocean, and geothermal energy, and conducts research on the construction of a low-carbon energy knowledge organization system within the multifaceted technology classification system.

For each of the seven types of energy, the main technology classifications, industries, and industry domains are thoroughly examined, with the corresponding IPC and CPC classification numbers documented. The attribute characteristics of these energies, as classified in different systems, are unified to facilitate the realization of conceptual and data level fusion, resulting in a comprehensive global system.

The knowledge organization process for wind energy is exemplified by its integration of a technology classification system that delineates wind energy and its associated technology branches. Initially, the technology classification system of wind energy is merged, providing a comprehensive overview of wind energy and its associated technology branches. From a textual feature perspective, the first level

encompasses all wind energy (F03D), adhering to the principle of majority same, ensuring a precise alignment between the two-ontology information. Subsequently, from a structural feature perspective, the subordinate technology branches, such as F03D1 with F03D as the parent node have a higher probability of being matched; third, H02J3/38 has a matching relationship with multiple entities belonging to the same class of H02J (wind power generation).

Knowledge Graph in the field of low-carbon and zero-carbon

Based on a multi-energy knowledge organization system, a multi-energy knowledge graph is built by collecting data, extracting attributes, aligning, building relationships, building concept hierarchies, and classifying entities. As shown in Fig.4 and Fig.5, entity, relationship and attribute knowledge are extracted from patent fields, and structured knowledge within the knowledge graph is linked and enhanced using knowledge graph construction techniques like entity linking and entity complementation. Finally, the knowledge graph in the field of low-carbon energy technologies will be formed and stored in the Neo4j.

Entity type	Entity Examples
Patent	US20130259187A1
	Rockfuel Innovations Limited
	Ipswich Suffolk GB
	C3583
	C10K
	6.2

Entities	Relations
1698	195,475

RelationTypes	Relation Examples
Coexist	(US9748006B2, COEXIST, US9691508B2)
IPCis	(FR920290A, C2619, 1, GMJJIS)
	(FR1180356A, 3.2.1, XXCYIS)
	(FR920290A, C01B,1, IPCIS)
Belongs	(FR920290A, ICI LTD, 1, BELONGS)
Whereis	(JP2905705B2, Japan, 1, WHEREIS)
	(SHINKO PANTETSUKU KK, Japan, 1, COUNTRYIS)

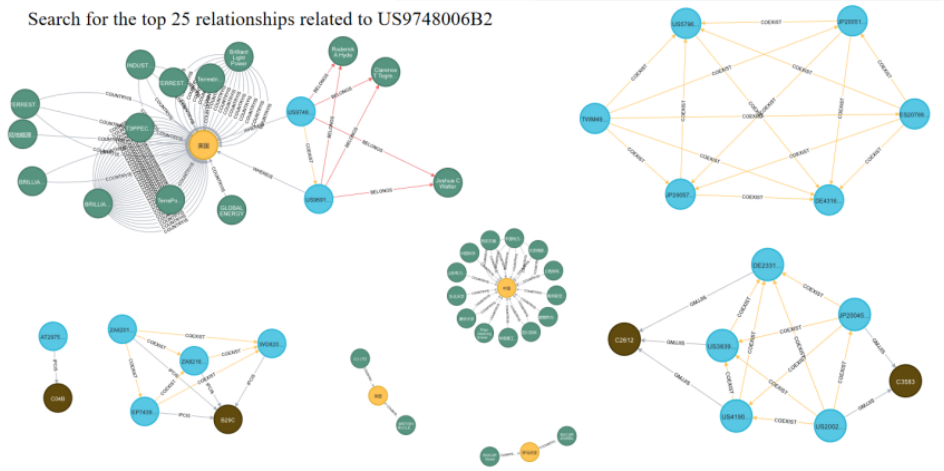


Figure 4. Example of Knowledge Graph Entity Relationships in the Low-Zero Carbon Domain.

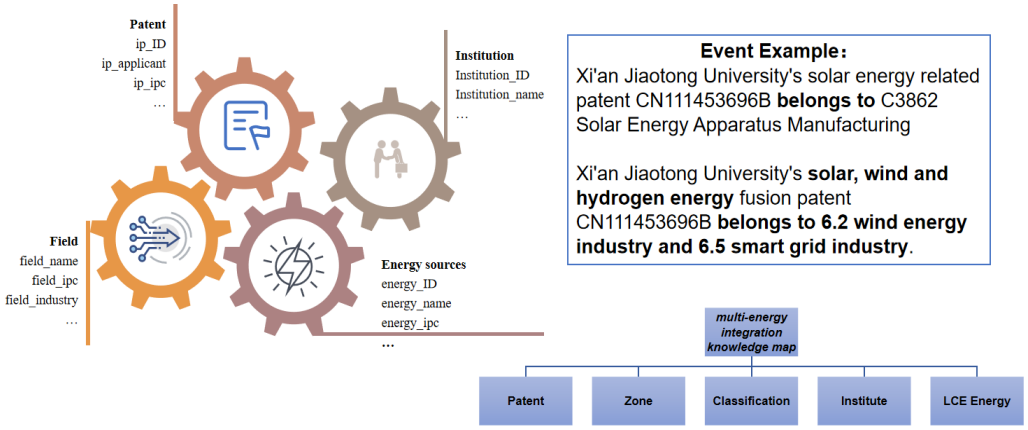


Figure 5. Event Example for Multi-energy Knowledge Graph.

Evolutionary Analysis of Technology Convergence Paths in the Low Carbon Energy Field

This study employs a knowledge discovery method based on the integration of multiple low-carbon and clean energy technology classifications; a methodology previously outlined in prior research. This paper utilizes patent analysis and application to examine the current state of knowledge reserves in the field, with the aim of identifying opportunities for technological innovation and evolutionary paths in the field of low-carbon and clean energy.

Seven low-carbon clean energy technology convergence trends

(1) The first signal of convergence is an inevitable product of the development of the multi-energy technology——from wind and ocean power to nuclear, hydrogen and solar PV. As it shown in Fig.6, in the 19th and early 20th centuries, hydroelectricity and wind power accounted for the largest share at over 60%. After the World War II there was a shift to nuclear fission, with nuclear power accounting for up to 46% of the total, and after 1975 a shift to solar PV and wind power, which together accounted for almost 45% of the total in the same period. The analysis revealed a general upward trend in renewable energy-related patents, with a marked increase occurring subsequent to 1980 and accelerated growth following 2005. The study also noted considerable variability in the development of distinct energy technologies, with wind energy patents dating back to 1907 and biomass-related patents emerging only after 1970. The patents demonstrating the most significant growth and proliferation are those associated with wind and solar energy. In contrast, patents related to ocean water energy, hydrogen energy, biomass energy, and nuclear energy have experienced a notable surge in innovation between 2011 and 2015, followed by a subsequent decline in recent years.

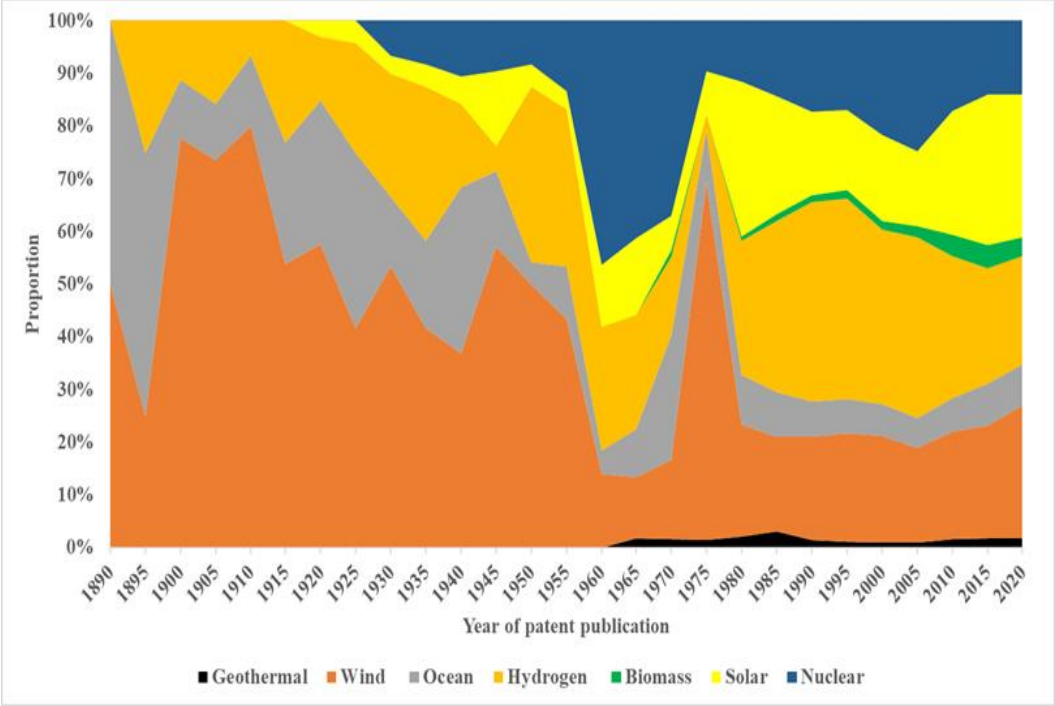


Figure 6. Trends in Multiple Energy Patent Applications (1890-2020).

(2) A second signal of technology convergence is the increasing share of cited papers in the total number of patents (Fig.7). The first patent literature on renewable energy technologies emerged in the early 20th century, yet the majority of citations to relevant scientific papers within this patent literature materialized subsequent to the 1970s. This observation suggests a growing reliance on scientific research with the progression of low-carbon clean energy technologies, particularly within the domain of biomass energy. Along with the steady rise in the number of patents filed for these seven low-carbon clean energy technologies, there has been an analogous increase in the proportion of relevant patents citing relevant papers. In recent years, biomass energy has become the most science-dependent energy source due to its close links to biochemical research, with 33-57% of cited papers. Nuclear, hydrogen and photovoltaics are also more science-dependent than other technologies (especially hydro and wind). Science-intensive technologies tend to be more dependent on basic science than on applied technologies across a wide range of energy sectors. A close examination of the seven low-carbon clean energy technologies reveals that the development of biomass energy technology is most dependent on basic scientific research. This is due to the fact that the technology is closely related to biochemical research. Consequently, this feature is the most distinctive. In contrast, the remaining energy technologies, particularly hydropower and wind energy, exhibit a stronger correlation with applied research.

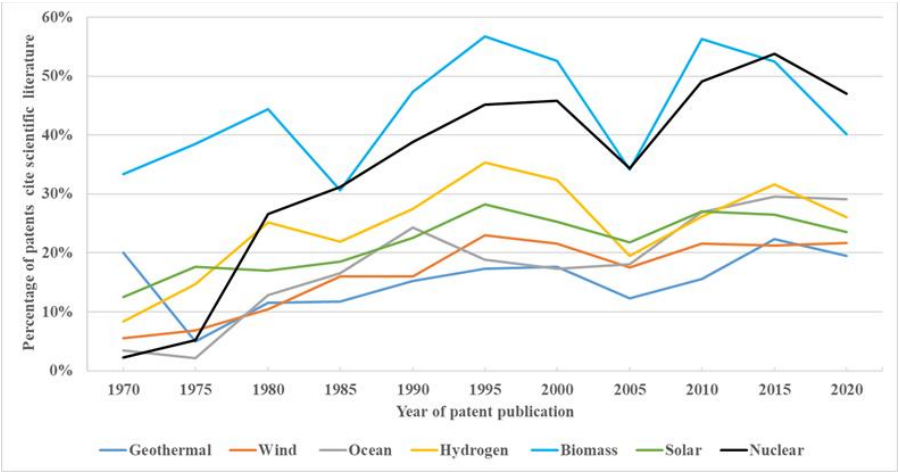


Figure 7. Trends in the percentage of patents citing paper (1970-2020).

(3) A third signal of technological convergence is that the convergence of multiple energy sources is based on the same or similar scientific principles. Statistics on the number and type of IPC top 4 for the full set of patents for various LCE show that 221 sub-categories are involved in the field, of which solar PV has the highest number of IPC categories involved with 81.The basis for the multi-energy integration of the various energy is the physics of nuclear energy integration based on plasma (G), ocean and geothermal energy based on mechanical engineering (F), biofuels and hydrogen fuel cells based on chemistry (C), and solar PV and wind energy based on photoelectric effects (F&H). Tables 3 and 4 provide mutual corroboration of the basis for the convergence of multiple energy technologies in terms of quantity and type, respectively.

Table 2. Number of IPC Top 4-Digit Classification Based on Seven Energy Sources.

Energies	G	B	C	E	F	H
Nuclear	**215**	98	269	2	84	62
Wind	584	602	176	466	7687	4418
Solar	632	557	706	420	**4827**	**4607**
Biomass	3	297	**1799**	2	183	164
Ocean	19	301	83	124	**4329**	249
Geothermal	9	16	20	81	**847**	66
Hydrogen	178	598	2640	31	902	737

Table 3. Types of IPC Top 4-Digit Classification Based on Seven Energy Sources.

Energies	G	B	C	E	F	H	Y02
Nuclear	**12**	2	1	2	1	0	3
Wind	0	6	1	2	**15**	8	6
Solar	1	1	2	2	**25**	**33**	17
Biomass	1	2	24	1	5	1	0
Ocean	0	2	1	4	9	1	2
Geothermal	0	0	0	2	12	2	1
Hydrogen	0	1	0	2	1	1	4

(4) The RTA was used to compare and analyze the patent-protected technologies of major global science and technology powers (China, Japan, the United States, Germany, and South Korea). These countries were selected based on their status in two key fields: low-carbon and zero-carbon energy, as well as the storage and convergence of multiple energy sources.

(5) The calculation results reveal that China possesses a substantial relative advantage in the domain of low-carbon clean energy (Fig.8), with its renewable energy patented technology. Japan's strengths lie primarily in nuclear power and non-electric utilization technology, while the United States leads in nuclear energy, hydrogen energy, and fuel cells. In the domain of energy storage and multi-energy integration (Fig.9), China has a substantial advantage in heat and cold storage, new power systems based on renewable energy, and chemical storage. In contrast, Japan and the United States prioritize chemical storage, while Germany focuses on physical storage.

(6) The world's major technological powers each possess a distinct set of advantages in terms of energy sources. The RTA indicator was utilized to calculate the relative advantages of seven types of patented technologies within the five major science and technology powerhouses. The results of this analysis indicate that: China has three types of energy with relative advantages in patented technologies (RTA $\geqslant$ 0.9), in the order of solar energy, wind energy, and geothermal energy. The United States has six types of advantageous energy technologies, in the order of biomass, hydrogen, nuclear energy, solar energy, geothermal energy, and wind energy, in which the biomass, hydrogen, nuclear energy, and patent advantages are also ahead of the other four countries. South Korea has four types of advantageous energy technologies, in the order of ocean water, solar energy, wind energy, and geothermal energy. The United States boasts six predominant energy technologies: biomass, hydrogen, nuclear, solar, geothermal, and wind. Among these, biomass, hydrogen, and nuclear lead the other four countries in terms of patent superiority.

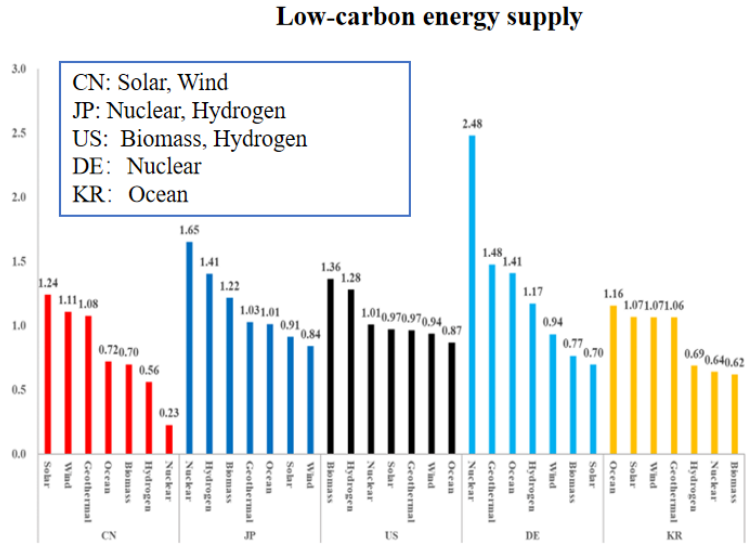


Figure 8. RTA of Low-carbon energy supply in key countries.

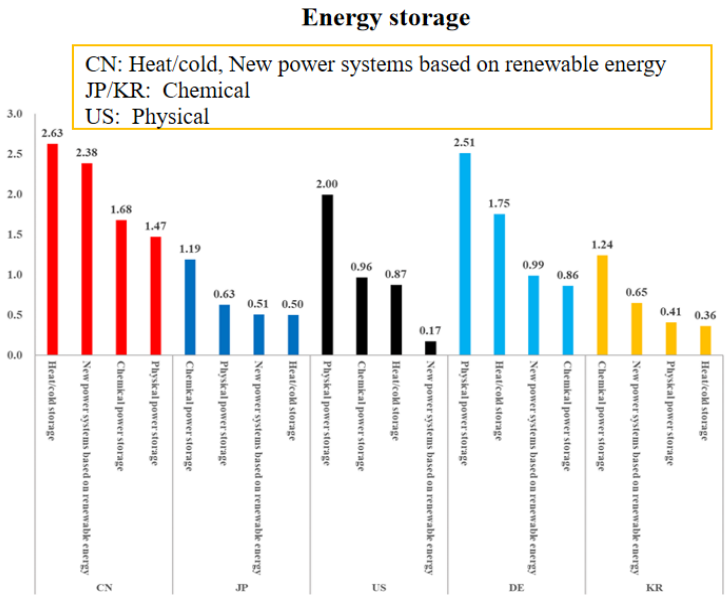


Figure 9. RTA of energy storage in key countries.

Multi-energy convergence potential detection

A patent-based technology convergence analysis has been conducted on seven significant domains of low-carbon clean energy technology: nuclear power and non-electric utilization of nuclear power, wind energy, solar energy, biomass energy, geothermal energy, ocean energy, hydrogen energy, and fuel cells. This analysis integrates the frequency of co-occurrence of technologies and the degree of technology relevance to identify key technologies, the degree of technology

convergence within the field, and the future development trends in the domain of low-carbon clean energy.

(1) The number of co-occurrences of the seven energy patents is 3.44% (Table 5). The seven energy sources have few connections. There is not enough technological correlation between two of the seven energy sources, but wind energy shows a weak correlation with geothermal energy, ocean energy, solar energy, geothermal energy with solar energy, and hydrogen energy with biomass energy. It is more obvious when wind is combined with other energy sources.

(2) By measuring the correlation coefficient S_{ij} for n core technologies, a new n*n diagonal matrix can be obtained to demonstrate the proximity of the integration between core technologies (Table 5).

As illustrated in Table 5, the low-carbon clean energy technology combinations that demonstrate a certain degree of correlation between the patented technologies include wind energy-geothermal energy, wind energy-ocean energy, wind energy-solar energy, geothermal energy-solar energy, and hydrogen energy-biomass energy. When these findings are considered in conjunction with the scale of the number of patents, they serve to further substantiate the significance of wind energy in the development of multi-energy technology integration.

(3) The study takes the knowledge map of wind energy integration with other energies as an example (Fig.10). Wind energy, based on the photovoltaic effect, has a high degree of fusion with solar energy, which is mainly used for wind power generation and propulsion, as well as combinations with geothermal energy, ocean energy and nuclear energy. One of the most prominent directions of integration is the generation and propulsion of wind energy, which is based on the generation of power by mechanical means. Since the regions that are rich in wind energy are also likely to be rich in solar and geothermal energy, the possibility of their fusion association is higher. There is also integration with nuclear power.

Table 4. The co-occurrence of seven types of energy patents.

	Nuclear	Wind	Solar	Biomass	Geothermal	Ocean	Hydrogen	Total	coexist number	% of coexist	Total number of field
Nuclear	-	18	143	2	8	3	588	762	740	0.56%	131903
Wind	18	-	9416	44	156	4555	790	14979	14273	8.41%	169721
Solar	143	9416	-	206	852	640	1560	12817	12014	2.86%	420490
Biomass	2	44	206	-	7	19	2308	2586	2485	2.71%	91742
Geothermal	8	156	852	7	-	132	55	1210	1067	8.65%	12333
Ocean	3	4555	640	19	132	-	305	5654	5239	13.78%	38011
Hydrogen	588	790	1560	2308	55	305	-	5606	5165	2.61%	198266

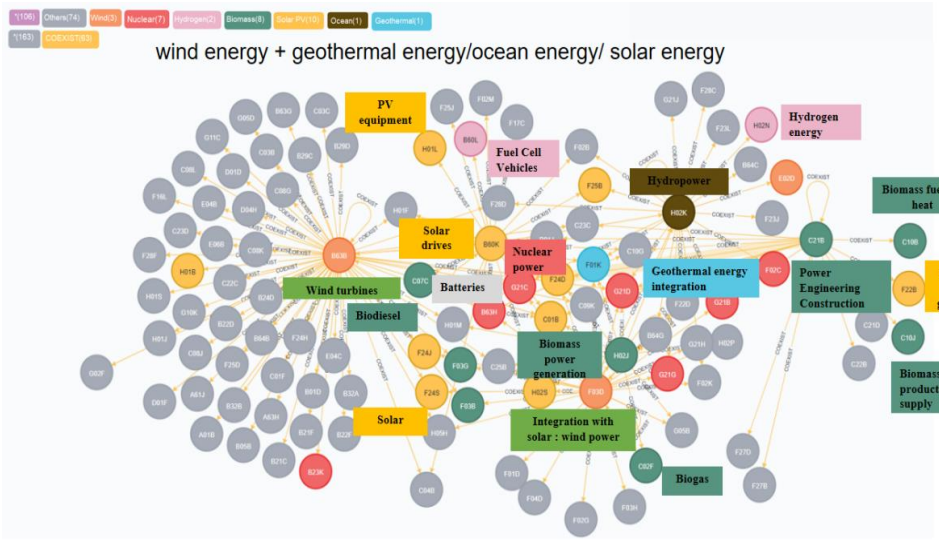


Figure 10. The case of Knowledge Graphs of wind and other energy convergence.

Technology Convergence Discovery and Evolution Analysis

Nuclear energy is a reliable, carbon-free energy source that generates a stable, continuous supply of electricity. In order to further analyse the evolutionary path of technological convergence, the convergence in the field of nuclear energy is selected as case study. The study is divided into five time periods: 1996-2000, 2001-2005, 2006-2010, 2011-2015, and 2016-2021. The co-occurrence network of convergence technologies in each period was extracted (Fig.11), the co-occurrence network of countries and technological field was analysed, and the development path of fusion technology was analysed. As shown in Figure 27, the number of patent applications in the field of nuclear energy has increased steadily with small fluctuations, peaking in 2004, 2012, and 2018, respectively. The number of IPC classification numbers related to nuclear fusion technology has been steadily increasing, indicating the continuous absorption of new technologies. Overall, the technology network was relatively isolated in its early years (1996-2000). However, as the 21st century progressed, the technical and functional network was gradually improved, and fusion trends began to emerge. The United States and Japan are leaders. In the last five years (2016-2021), China has joined in, forming a triad.

The United States and Japan have a clear first-mover advantage in the development of patented technology convergence in the field of nuclear energy. Since 2011, China has witnessed a substantial surge in the number of patents related to nuclear energy technology, which have begun to occupy a central position in the evolutionary network of fusion technology. In the past five years (2016-2021), China, the United States, and Japan have further solidified their dominance in this field. When considering China's historical context of related planning, the role of policy as an incentive for innovation becomes evident. Since 2005, China's "Eleventh Five-Year Plan" has promoted the development of nuclear power policy from "moderate

development" to "active development." In 2006, China initiated the process of third-generation nuclear power autonomy. In 2011, China introduced the "Medium- and Long-Term Development Plan for Nuclear Power (2011-2020)," which adjusted the development goal to 58 million kilowatts of installed nuclear power in operation by 2020, with 30 million kilowatts of nuclear power under construction.

As illustrated in Fig. 11, the progression of patented nuclear energy technology fusion follows a technological trajectory that primarily involves the conversion of energy between nuclear, thermal, mechanical, and electrical domains. In terms of technology categories, batteries and their manufacture (H01M and its subordinate branches), chemical or physical methods and devices, such as catalysis, colloid chemistry (B01J and its subordinate branches), have been almost throughout the entire process of nuclear energy fusion technology evolution, suggesting that they are the basic and key technologies in the field of nuclear energy technology fusion. Metal compounds (C01G), electrolytic processes to produce compounds or non-metals (C25B), alloys (C22C), engines (F03G), ion implantation or chemical vapor deposition (C23C), and so forth, have played a significant and innovative role in the fusion of nuclear energy and other energy sources at various points in time, propelling technological turnover. This demonstrates that nuclear energy's function extends beyond mere electricity supply, encompassing the production of hydrogen, district heating, desalination, and numerous other nuclear technologies. It demonstrates that nuclear energy's function extends beyond the provision of electricity, encompassing diverse non-electricity-related applications such as hydrogen production, district heating, and seawater desalination.

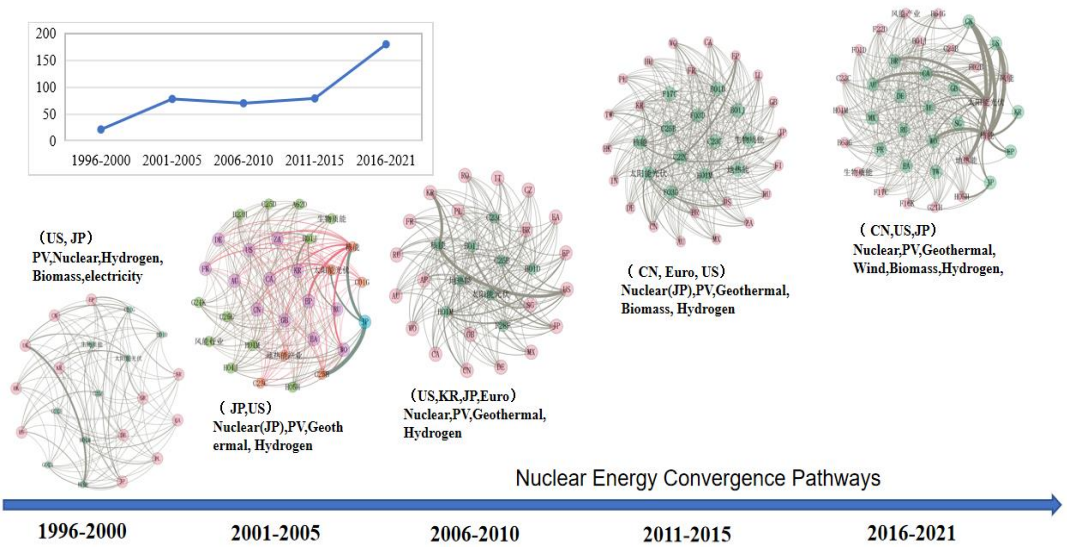


Figure 11. The Nuclear Energy Convergence Pathways.

In essence, the primary characteristics of the development of nuclear energy technology integration and innovation path performance can be delineated as follows: In the initial phase, the emphasis was placed on the augmentation of primary energy, with nuclear energy serving as a reliable power source. The technological innovation agenda centered on nuclear energy and renewable energy coupling technologies and facilities, including, but not limited to, small nuclear reactors and centralized solar thermal power plant technology. In recent times, the emphasis has shifted towards enhancing energy efficiency, with a particular focus on nuclear energy and renewable energy synergistic smart systems. Technological innovation has centered on the power system as the core, leveraging smart grids, with nuclear energy and renewable energy serving as the primary sources, complemented by an appropriate amount of hydropower and thermal power. This approach aims to facilitate the complementarity of cold, heat, gas, water, electricity, and other energy sources, thereby enhancing the efficiency of energy utilization. (iii) The present focus is on the expansion of the application of multi-energy fusion technology, with attention given to technological innovations related to nuclear energy for hydrogen production, such as large-scale nuclear energy for hydrogen production and secondary energy production.

Discussion & Conclusions

This paper proposes a domain knowledge discovery method based on convergence perspective, which can reveal domain knowledge reserve dynamics, technology opportunity insight and domain convergence technology evolution path. The method can help researchers, enterprises and government departments to better understand the technological opportunities, key technologies and future development trends of field convergence, which is important for promoting innovation and development of related fields.

The Multi-Energy Convergence Patent Knowledge Graph (MEPKG) is a domain knowledge database that extracts the latest advances in domain knowledge and provides information on the potential for technology convergence and development trajectories. It also improves user search experience, search engines, and knowledge discovery. In this paper, based on the research issues of multi-energy convergence (research progress, convergence potentials, and convergence paths), we try to use the knowledge graph of multi-energy convergence patents to study the trends, convergence signals, potentials, and development paths of multi-energy integration. The empirical analysis reveals that the signals of technology convergence in the multi-energy field are relatively weak. At present, the technological links among the seven energy sources are not very significant, but the integration of wind energy with other energy sources is more obvious. Due to the fusion effect of multiple energy sources between scientific principles, its potential for future multi-energy convergence is huge. In terms of technology pathways, the focus is on multi-energy

power generation and thermal efficiency utilization. Patent-based multi-energy knowledge mapping helps to detect weak signals of multi-energy technology convergence. The study of technology pathways for their convergence is currently dominated by multi-energy power generation and thermal efficiency utilization. It can also reveal the evolutionary path of convergence of individual energy sources.

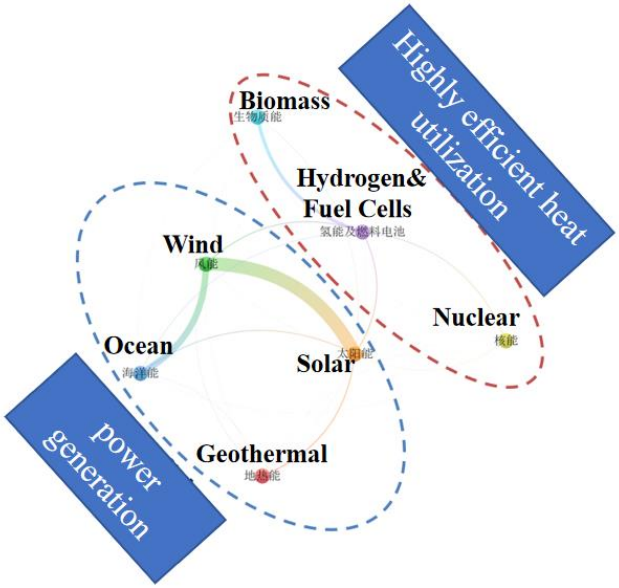


Figure 12. Multi-energy convergence directions.

However, the limitation of this paper is that the current research method is still based on structured data of patents. It is based on the fusion of the patent classification numbers, which belongs to a relatively elementary stage of the attempt. In the future, structured data and unstructured heterogeneous data will be further explored to further improve the research and application of knowledge graph on patent technology analysis. The knowledge mapping technology has great expansion potential in multi-domain and cross-domain to further improve data availability and data processing capability. Second, the boundaries between multi-domain and cross-domain are evolving with the dynamic integration of cross-domain, and future attention will be paid to the detection of opportunities for the integration of emerging technologies. In the future, we will continue to refine the method, expand the application areas, and improve the accuracy and efficiency of domain knowledge discovery. At the same time, we hope that the method will provide more valuable support and assistance to research and development in related fields.

There are several areas that require further exploration through research. Firstly, the automatic mapping model of multivariate technology classification categories necessitates the refinement of algorithms related to matching rules. Secondly, the indicators and algorithms for technology convergence discovery and path evolution analysis require enhancement, and the correlation with the knowledge graph must be strengthened. In future research, the application of the knowledge map of low-carbon

and clean energy can be further strengthened, and the algorithms related to technology fusion path and evolution analysis can be improved through the use of algorithms related to the graph database.

Acknowledgments

The research is the outcome of the projects, "Youth Innovation Promotion Association (2022173)" "Light of West China program", "Intelligence Service for Research Institutes on Science, Technology and Innovation (E3291106)", "Research on Emerging Technology Direction Identification Method based on Technology Fusion (E3Z0000803)", supported by Chinese Academy of Sciences(CAS), "Study on the Multi-Relation Data Fusion Methods for Identification and Prediction of Technology Innovation Paths"(No. 18BTQ067) supported by National Social Science Fund of China and "Early Recognition Method of Transformative Scientific and Technological Innovation Topics based on Weak Signal Temporal Network Evolution analysis" (No.72274113) supported by the National Natural Science Foundation of China. Also, the special contribution of Chunjiang Liu as C.Author liucj@clas.ac.cn.

Reference

Agrawal, R. and Shafer, J.C. (1996) 'Parallel mining of association rules', *IEEE Transactions on Knowledge and Data Engineering*, 8(6), pp. 962–969. Available at: https://doi.org/10.1109/69.553164.

Albino, V. et al. (2014) 'Understanding the development trends of low-carbon energy technologies: A patent analysis', Applied Energy, 135, pp. 836–854. Available at: https://doi.org/10.1016/j.apenergy.2014.08.012.

Anguera, A. et al. (2016) 'Applying data mining techniques to medical time series: an empirical case study in electroencephalography and stabilometry', *Computational and Structural Biotechnology Journal*, 14, pp. 185–199. Available at: https://doi.org/10.1016/j.csbj.2016.05.002.

Balaid, A. et al. (2016) 'Knowledge maps: A systematic literature review and directions for future research', *International Journal of Information Management*, 36(3), pp. 451–475. Available at: https://doi.org/10.1016/j.ijinfomgt.2016.02.005.

Chen, Y.-H., Chen, C.-Y. and Lee, S.-C. (2011) 'Technology forecasting and patent strategy of hydrogen energy and fuel cell technologies', *International Journal of Hydrogen Energy*, 36(12), pp. 6957–6969. Available at: https://doi.org/10.1016/j.ijhydene.2011.03.063.

CNIPA (2023) Patent Classification System for Green and Low Carbon Technologies. Available at: https://www.cnipa.gov.cn/module/download/downfile.jsp?classid=0&showname=%E7%BB%BF%E8%89%B2%E4%BD%8E%E7%A2%B3%E6%8A%80%E6%9C%AF%E4%B8%93%E5%88%A9%E5%88%86%E7%B1%BB%E4%BD%93%E7%B3%BB.pdf&filename=cda44a0d91494f879710e61f9f8112a2.pdf.

Curran, C.-S., Bröring, S. and Leker, J. (2010) 'Anticipating converging industries using publicly available data', *Technological Forecasting and Social Change*, 77(3), pp. 385–395. Available at: https://doi.org/10.1016/j.techfore.2009.10.002.

Dessì, D. et al. (2021) 'Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain', *Future*

Generation Computer Systems, 116, pp. 253–264. Available at: https://doi.org/10.1016/j.future.2020.10.026.
DOE (2022) Nuclear Hydrogen R&D Plan. Available at: https://www.energy.gov/sites/prod/files/2015/01/f19/fcto_nucle ar_h2_r%26d_plan.pdf.
Fayyad, U. (2001) 'Knowledge Discovery in Databases: An Overview', in S. Džeroski and N. Lavrač (eds) Relational Data Mining. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 28–47. Available at: https://doi.org/10.1007/978-3-662-04599-2_2.
Fu, P. et al. (2020) 'Integration of Hydrogen into Multi-Energy Systems Optimisation', Energies, 13(7), p. 1606. Available at: https://doi.org/10.3390/en13071606.
Gregory Piatetsky-Shapiro (1991) 'Report on the AAAI-91 Workshop on Knowledge Discovery in Databases', *IEEE Intelligent Systems & Their Applicatiobns*, 6(5), pp. 74–76.
Guo, R., Lv, S., Liao, T., Xi, F., Zhang, J., Zuo, X., Cao, X., Feng, Z., & Zhang, Y. (2020). 'Classifying green technologies for sustainable innovation and investment. Resources', *Conservation and Recycling*, 153, 104580.
Hötte, K., Pichler, A. and Lafond, F. (2021) 'The rise of science in low-carbon energy technologies', *Renewable and Sustainable Energy Reviews*, 139, p. 110654. Available at: https://doi.org/10.1016/j.rser.2020.110654.
Jeon, J. and Suh, Y. (2019) 'Multiple patent network analysis for identifying safety technology convergence', *Data Technologies and Applications*, 53(3), pp. 269–285. Available at: https://doi.org/10.1108/DTA-09-2018-0077.
Jiang, L. et al. (2022) 'Patent analysis for generating the technology landscape and competition situation of renewable energy', *Journal of Cleaner Production*, 378, p. 134264. Available at: https://doi.org/10.1016/j.jclepro.2022.134264.
JPO (2022) Green Transformation Technologies Inventory. Available at: https://www.meti.go.jp/press/2022/06/20220623001/20220623001.html.
Karlapalem, K. (ed.) (2021) Advances in knowledge discovery and data mining: 25th Pacific-Asia conference, PAKDD 2021, virtual event, May 11-14, 2021: *proceedings. Part 1. PAKDD, Cham: Springer* (Lecture notes in computer science Lecture notes in artificial intelligence, 12712).
Kim, J. and Lee, S. (2017) 'Forecasting and identifying multi-technology convergence based on patent data: the case of IT and BT industries in 2020', *Scientometrics*, 111(1), pp. 47–65. Available at: https://doi.org/10.1007/s11192-017-2275-4.
Lee, C., Kogler, D.F. and Lee, D. (2019) 'Capturing information on technology convergence, international collaboration, and knowledge flow from patent documents: A case of information and communication technology', *Information Processing & Management*, 56(4), pp. 1576–1591. Available at: https://doi.org/10.1016/j.ipm.2018.09.007.
Lee, C., Park, G. and Kang, J. (2018) 'The impact of convergence between science and technology on innovation', *The Journal of Technology Transfer*, 43(2), pp. 522–544. Available at: https://doi.org/10.1007/s10961-016-9480-9.
Lee, J. and Fink, D. (2013) 'Knowledge mapping: encouragements and impediments to adoption', *Journal of Knowledge Management*, 17(1), pp. 16–28. Available at: https://doi.org/10.1108/13673271311300714.
Li, W. et al. (2022) 'An Approach to Achieve Carbon Neutrality with Integrated Multi-Energy Technology', *Engineering*, 19, pp. 11–13. Available at: https://doi.org/10.1016/j.eng.2021.09.024.
Li, Z., He, T. and Farjam, H. (2023) 'Application of an intelligent method for hydrogen-based energy hub in multiple energy markets', *International Journal of Hydrogen Energy*, 48(93), pp. 36485–36499. Available at:

https://doi.org/10.1016/j.ijhydene.2023.03.124.
Liu, J.S. et al. (2011) 'Photovoltaic technology development: A perspective from patent growth analysis', *Solar Energy Materials and Solar Cells*, 95(11), pp. 3130–3136. Available at: https://doi.org/10.1016/j.solmat.2011.07.002.
Mariscal, G., Marbán, Ó. and Fernández, C. (2010) 'A survey of data mining and knowledge discovery process models and methodologies', *The Knowledge Engineering Review*, 25(2), pp. 137–166. Available at: https://doi.org/10.1017/S0269888910000032.
National Nuclear Laboratory (2021) Unlocking the UK's Nuclear Hydrogen Economy to Support Net Zero. Available at: https://www.nnl.co.uk/wp-content/uploads/2021/07/Hydrogen-Round-Table-FINAL-v2.pdf.
Nguyen, S.H. and Chowdhury, G. (2013) 'Interpreting the knowledge map of digital library research (1990–2010)', *Journal of the American Society for Information Science and Technology*, 64(6), pp. 1235–1258. Available at: https://doi.org/10.1002/asi.22830.
Oltra, V. and Saint Jean, M. (2009) 'Variety of technological trajectories in low emission vehicles (LEVs): A patent data analysis', *Journal of Cleaner Production*, 17(2), pp. 201–213. Available at: https://doi.org/10.1016/j.jclepro.2008.04.023.
Park, I. and Yoon, B. (2018) 'Technological opportunity discovery for technological convergence based on the prediction of technology knowledge flow in a citation network', *Journal of Informetrics*, 12(4), pp. 1199–1222. Available at: https://doi.org/10.1016/j.joi.2018.09.007.
Pino-Díaz, J. et al. (2012) 'Strategic knowledge maps of the techno-scientific network', *Journal of the American Society for Information Science and Technology*, 63(4), pp. 796–804. Available at: https://doi.org/10.1002/asi.21712.
Roscher, R. et al. (2020) 'Explainable Machine Learning for Scientific Insights and Discoveries', IEEE Access, 8, pp. 42200–42216. Available at: https://doi.org/10.1109/ACCESS.2020.2976199.
Sebastian, Y. and Then, P.H.H. (2011) 'Domain-driven KDD for mining functionally novel rules and linking disjoint medical hypotheses', *Knowledge-Based Systems*, 24(5), pp. 609–620. Available at: https://doi.org/10.1016/j.knosys.2011.01.008.
Su, H.-N. and Lee, P.-C. (2010) 'Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight', *Scientometrics*, 85(1), pp. 65–79. Available at: https://doi.org/10.1007/s11192-010-0259-8.
Subtil Lacerda, J. (2019) 'Linking scientific knowledge and technological change: Lessons from wind turbine evolution and innovation', *Energy Research & Social Science*, 50, pp. 92–105. Available at: https://doi.org/10.1016/j.erss.2018.11.012.
Suresh, R.H. and Egbu, C.O. (2004) 'KNOWLEDGE MAPPING: CONCEPTS AND BENEFITS FOR A SUSTAINABLE URBAN ENVIRONMENT', in. 20th Annual ARCOM Conference, eriot Watt University.
USPTO (2009) Environmentally Sound Technologies (EST) concordance. Available at: https://www.uspto.gov/web/patents/classification/international/est_concordance.htm.
USPTO；EPO (2010) COOPERATIVE PATENT CLASSIFICATION. Available at: https://www.uspto.gov/web/patents/classification/cpc/html/cpc-Y02E.html#Y02E.
Vickery, B. (1997) 'Knowledge discovery from databases: an introductory review', Journal of Documentation, 53(2), pp. 107–122. Available at: https://doi.org/10.1108/EUM0000000007195.
WIPO (2010) WIPO IPC Green Inventory. Available at: https://www.wipo.int/classifications/ipc/green-inventory/home.

Xiao, Y., Li, C. and Thürer, M. (2023) 'A patent recommendation method based on KG representation learning', *Engineering Applications of Artificial Intelligence*, 126, p. 106722. Available at: https://doi.org/10.1016/j.engappai.2023.106722.

Xue, D. and Shao, Z. (2024) 'Patent text mining based hydrogen energy technology evolution path identification', *International Journal of Hydrogen Energy*, 49, pp. 699–710. Available at: https://doi.org/10.1016/j.ijhydene.2023.10.316.

Yamasaki, S., Yaji, K. and Fujita, K. (2019) 'Knowledge discovery in databases for determining formulation in topology optimization', *Structural and Multidisciplinary Optimization*, 59(2), pp. 595–611. Available at: https://doi.org/10.1007/s00158-018-2086-0.

Yue, M. et al. (2021) 'Hydrogen energy systems: A critical review of technologies, applications, trends and challenges', *Renewable and Sustainable Energy Reviews*, 146, p. 111180. Available at: https://doi.org/10.1016/j.rser.2021.111180.

Zhou, H. et al. (2024) 'Understanding innovation of new energy industry: Observing development trend and evolution of hydrogen fuel cell based on patent mining', *International Journal of Hydrogen Energy*, 52, pp. 548–560. Available at: https://doi.org/10.1016/j.ijhydene.2023.07.032.

https://doi.org/10.51408/issi2025_092

Exploring Nobel Laureates' Question Selection Characteristics from a Topical Perspective*

Zou Xinran[1], Dong Yu[2], Wu Jiaxin[3]

[1]*zouxinran@mail.las.ac.cn*, [2]*dongy@mail.las.ac.cn*
National Science Library, Chinese Academy of Sciences, Beijing (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing (China)

[3]*shanganchong@163.com*
Department of Management Science and Engineering, School of Economics and Management, Wuhan University, Wuhan (China)

Abstract

Selecting a research question is the starting point of scientists' research activities, playing a crucial role not only in their career development, and further exploration is needed in this area. In this paper, we construct a characteristic index system for scientists' question selection based on the quality of attention theory in psychology, then conduct an empirical analysis using Nobel Laureates in natural sciences as examples, to reveal the characteristics of their question selection. Results show that Nobel laureates exhibit both commonalities and disciplinary differences in their question selection. Common characteristics include: a concentration of research topics in a limited number of directions, strong persistence in their research focus, and a balanced allocation of research effort between Broad and focus. Disciplinary differences are also evident. Physics laureates tend to engage in sustained and steady research across multiple interrelated fields. Chemistry laureates show a relatively higher degree of cross-disciplinary and cross-domain question selection; while they may moderately shift research directions over the course of their careers, these shifts typically revolve around one or a few core themes. In contrast, laureates in Physiology or Medicine display more exploratory question selection behaviors, frequently switching among one or several related core areas, with comparatively lower research persistence. Across different stages of their careers, the three groups of laureates demonstrate distinct question selection patterns. Physics laureates tend to broaden the scope of their research while simultaneously deepening their focus. Chemistry and Physiology or Medicine laureates follow a similar trajectory characterized by early-stage broad exploration, mid-career flexibility, and late-career deep focus. These findings highlight the varying research patterns across disciplines and offer valuable insights into how Nobel laureates select and shift their research questions.

Introduction

Selecting a research question is the starting point of scientists' research activities, playing a crucial role not only in their career development but also in shaping the progress of their discipline (Ding et al., 2023; Yu et al., 2021; Foster et al., 2015).

* This work is supported by the Major Project of the National Social Science Foundation of China, "Research on Talent Training Models and Strategic Deployment in Key and Core Technology Fields in China" (22&ZD127).

Exploring the characteristics of research question selection provides valuable insights into scientists' career development process and contributes to a deeper understanding of their research patterns. Moreover, it holds important practical implications for science and technology management, particularly in areas such as talent cultivation, research funding allocation, and disciplinary development.

The selection of research questions by scientists has long been a central topic of interest in the fields of sociology and scientometrics (Ding et al., 2023; Van Houten et al., 1983). As early as the 1980s, scholars began to investigate field mobility among physicists. Early studies were primarily based on qualitative research and simple questionnaire surveys. In recent years, the development of large-scale bibliometric datasets—such as Scopus, Microsoft Academic Graph, and the American Physical Society (APS) dataset—has provided rich data sources for quantitatively analyzing scientists' question selection. Moreover, the emergence of advanced techniques such as natural language processing and complex network analysis has offered new perspectives for exploring scientists' research behavior in greater depth (Liu et al., 2023; Taylor et al., 2022).

Current research on scientists' question selection has produced rich findings, with a primary focus on their question selection performance. However, the construction of measurement indicators related to the question selection performance often lacks theoretical grounding, resulting in a degree of subjectivity and the absence of a comprehensive and objective evaluation framework. Based on this, we introduce the attention theory in psychology and construct a characteristic index system of scientists' question selection. Then we conduct an empirical analysis using Nobel Laureates in natural sciences as examples. This study aims to enhance existing research and provide a more scientific basis and reference for the career development, talent training, and research funding policies of Chinese scientists.

Literature review

Analysis process and method of scientists' question selection characteristics

Existing studies primarily reveal the question selection characteristics of scientists by measuring the topic transition in their careers. The measurement process follows three main steps: (1) identifying research topics, (2) dividing time periods, and (3) measuring topic shifts.

There are three ways to identify the research topics of scientists. Many studies obtain the topic vector of a paper by vectorizing the topic code or research field assigned to the paper by the database. For example, the PACS code in the American Physical Society (APS) dataset consists of six letters and numbers, in which the first two numbers define 67 major topics in the field, and the topic vectors of multiple papers can be obtained by statistically normalizing the frequency of the first two numbers (Jia et al., 2017). The second is to construct citation or co-occurrence networks of papers and perform community clustering so as to classify each paper under different

subject categories. One study uses papers as nodes to build a paper co-citation network for each scientist. That is, if two papers cite the same references, they are linked together, and then use the Fast Unfolding algorithm to identify each scientist's co-citation network. The primary communities identified represent the scientist's main research topics (Zeng et al., 2019). Thirdly, the research topics can be identified from the title, keywords, abstract and other text information of the paper. For example, Bert (Ding & Chen, 2023), Top2Vec (Chen et al., 2019) are used to vectorize the text information such as the title of the paper, obtain the topic vector of each paper, and further determine the topic category of the paper through clustering. The evolution of a scientist's research topics is a dynamic process, and analyzing their entire career as a whole may overlook important temporal variations. To address this, existing studies segment a scientist's career into different time stages based on three main approaches: The fixed time interval approach groups papers published within a set number of consecutive years into the same stage. For instance, papers published in 2000 and 2001 are classified as one stage, while those from 2001 and 2002 form the next, and so on (Huang et al., 2023). The publication count approach segments a scientist's career by grouping a fixed number of consecutive papers into the same stage, regardless of the publication year. For example, some studies define each stage as a block of m consecutive papers (Huang et al., 2023). The key career milestone approach divides a scientist's career based on significant events such as the publication of their most cited paper or receiving an award. For instance, some researchers classify Nobel laureates' careers into three stages: before publishing their prize-winning paper, after publishing it, and post-award (Ding & Chen, 2023).

Existing research measures the characteristics of scientists' question selection across several dimensions: topic transition speed, topic transition span, topic focus intensity, and topic coverage. Topic transition speed measures how quickly scientists shift between different topics, reflecting their level of concentration on specific research problems. It can be quantified by calculating the amplitude of topic change over time or by the ratio of topics to papers within a given period (Chen et al., 2019). Topic transition span assesses the degree of content difference before and after a topic shift, indicating whether scientists tend to explore significantly new research directions. This is typically measured using cosine similarity (Liu et al., 2024) or Euclidean distance (Liu & Xia, 2017). Topic focus intensity captures the level of attention scientists devote to a particular topic within a specific timeframe, revealing whether they prefer deep exploration of a single issue. It is commonly measured by the number of papers published on a given topic—where a higher count indicates greater investment (Ding & Chen, 2023), or by the proportion of papers on a topic relative to the scientist's total output (Chen et al. 2023; Chen et al., 2019). Topic coverage reflects the breadth and diversity of a scientist's research within a given period. It can be measured by the number of topics studied (Ding & Chen, 2023), or spatially by calculating the coverage area of topic vectors. For example, the volume of the smallest ellipsoid encompassing all topic vectors can serve as a proxy (Bu et al., 2022).

Analysis subjects and conclusions of scientists' question selection characteristics

Most researches primarily focuses on scientists in physics and computer science. Studies on physicists often use the APS dataset, which assigns PACS codes to papers for topic identification. Research on computer scientists relies on multiple databases, including Microsoft Academic Graph (MAG), Microsoft Academic Search, and DBLP, all of which classify papers by research field. From a group perspective, some studies analyze a single cohort, such as all physicists or Nobel laureates in physics, while others compare multiple groups to examine differences in question selection。 Findings reveal some differences between elite scientists and the broader scientific. community. In physics, general scientists tend to expand their research over time: increasing the speed (Zeng et al., 2019), span (Aleta et al., 2019) and topic coverage (Zeng et al., 2019). In contrast, Nobel laureates remain more focused, dedicating long-term attention to their prize-winning topics and later expanding on related topics (Ding & Chen, 2023). A similar pattern is observed in computer science, where high-impact, high-productivity scientists exhibit greater research focus compared to others (Liu et al., 2024). Additionally, question selection characteristics vary across disciplines. For instance, scientists in physics are more likely to work on multiple topics simultaneously in the middle of their careers (Zeng et al., 2019), while scientists in computing are more likely to work early and late in their careers (Chakraborty et al., 2015). Although differences in research methodology and indicator selection may limit cross-disciplinary comparisons, the findings still highlight that scientists' question selection characteristics vary across disciplines.

Research Design

Topic identification

To measure the characteristic of scientists' question selection, it is essential to first identify the research topics of each scientist. We extract the research topics of each scientist from the collected paper datasets. Through text embedding, dimensionality reduction and topic clustering, the number of research topics for each scientist, the topic vectors of the research topics, and the topic classification of each paper are obtained.

Text Embedding

The title, abstract, and keywords of a paper condense the main research content of the paper and are a refined representation of the paper's topics. Text embedding can map the title, abstract, keywords and other information of the paper into vectors in the space, and extract the topic vector of the paper as the basis for subsequent topic clustering. SciBERT is a pre-trained language model based on the BERT architecture. Trained on a large-scale corpus of scientific publications, SciBERT offers stronger language understanding and semantic representation capabilities for scientific texts compared to other pre-trained models. Therefore, we use SciBERT to extract the topic vector of each paper.

Dimensionality reduction

In high-dimensional space, the distances among samples can become strikingly similar, making it difficult for clustering algorithms to distinguish distinct data characteristics. To alleviate this problem, we use the UMAP to reduce the dimensions of the vectors output by SciBERT.

Topic Clustering

AP (Affinity Propagation) clustering is suitable for small to medium-sized datasets and does not require a predefined number of clusters. Since most authors have fewer than 800 publications, this study applies the AP clustering algorithm to cluster the topic vectors of each scientist's papers. The clustering process yields the number of research topics for each scientist, assigns a topic label to each paper, and computes the average vector of papers within each cluster as the representation of that topic. The clustering results are evaluated using the Silhouette Coefficient (SC), which ranges from -1 to 1, with higher values indicating better clustering quality. Based on the SC, the number of research topics for each scientist is determined.

Research theory and Index construction

Psychological research widely recognizes that attention consists of four fundamental dimensions (Meng, 1994): attentional span, referring to the number of objects one can focus on simultaneously; attentional stability, denoting the ability to sustain attention on a specific perception or activity over time; attentional allocation, indicating the capacity to distribute attention across multiple objects or activities at the same time; and attentional shifting, describing the active, purposeful, and timely transition of attention from one object or activity to another.

Scientist's research question selection can be regarded as the allocation of research attention across different topics. Since a scientist's research attention is limited, they may adopt different allocation strategies, resulting in diverse patterns of question selection. Based on this, we refer to the attention quality theory described above to constructs an index system from four dimensions: span, stability, distribution and transfer.

Span dimension

Scientists may engage in multiple research topics within a given period. The degree of content variation among these topics reflects the breadth and diversity of their question selection. Accordingly, **Topic Coverage Index (*C*)** is introduced to measure the extent of content differentiation across a scientist's research topics during a specific time period. The formula is as shown in Eq. (1):

$$C = 1 - \min_{1 \le i < j \le N} s_{i,j} \tag{1}$$

$s_{i,j}$ represents the cosine similarity between v_i and v_j, calculated as shown in Eq. (2):

$$s_{i,j} = \frac{v_i \cdot v_j}{\|v_i\| \cdot \|v_j\|} \tag{2}$$

Where, v represents the topic vector of a scientists, and N represents the total number of research topics of the scientist.

Allocation dimension

When scientists engage in multiple research topics simultaneously, the amount of research effort devoted to each topic may vary. This can be measured by the number of publications under each topic—more publications in a given topic indicate a greater allocation of research effort to that area. An uneven distribution of publications across topics suggests the presence of core research areas, while a more balanced distribution indicates that the scientist tends to allocate research efforts more evenly across topics. Based on this, the **Topic Focus Index (F)** is introduced to reflect the degree of evenness in a scientist's allocation of research effort across different topics. The calculation is based on Pielou's Evenness Index, as shown in Eq. (3):

$$F = \frac{H'}{\log(N)} \tag{3}$$

H' is Shannon entropy, which can be calculated as shown in Eq. (4):

$$H' = -\sum_{i=1}^{N} p_i \log(p_i) \tag{4}$$

Where, p_i is the proportion of the number of papers under the i topic to the total number of papers, and the calculation formula is in Eq. (5)

$$p_i = \frac{n_i}{n} \tag{5}$$

The value of F ranges from 0 to 1, where a value closer to 1 indicates a more even distribution of a scientist's papers across topics. Here, n_i represents the number of papers published under topic i, n denotes the total number of papers published by a scientist, and N represents the total number of research topics explored by the scientist.

Stability dimension

Scientists may continue working on the same research topic over an extended period. The duration of sustained engagement with a topic reflects the persistence and stability of their research. The **Topic Duration Index (G)** is introduced to measure the proportion of a scientist's career spent on each topic, averaged across all topics. The formula is in Eq. (6):

$$G = \frac{1}{T \cdot N} \sum_{i=1}^{N} (t_i + 1)$$

(6)

Where t_i represents the difference between the publication year of the last and first paper under the topic i, N represents the total number of topics studied by the scientist, and T represents the span of a scientist's research career.

Shifting dimension

Scientists may shift research topics throughout their careers. The **Topic Shifting Speed Index (S)** is introduced to measure the frequency of transitions between different research topics over time. The formula is in Eq. (7):

$$S = \frac{1}{T-1} \sum_{i=2}^{T} N_i \quad (7)$$

Where T represents the number of years in which a scientist has published papers, i denotes a specific publication year, and N_i represents the number of new topics introduced in year i compared to year i-1. For example, if a scientist studied topic1, topic2, topic3 in year i-1, and in year i studied topic 1, topic2, topic4 and topic5, then N_i would be 2.

Empirical Study

Figure 1 shows the technology roadmap.

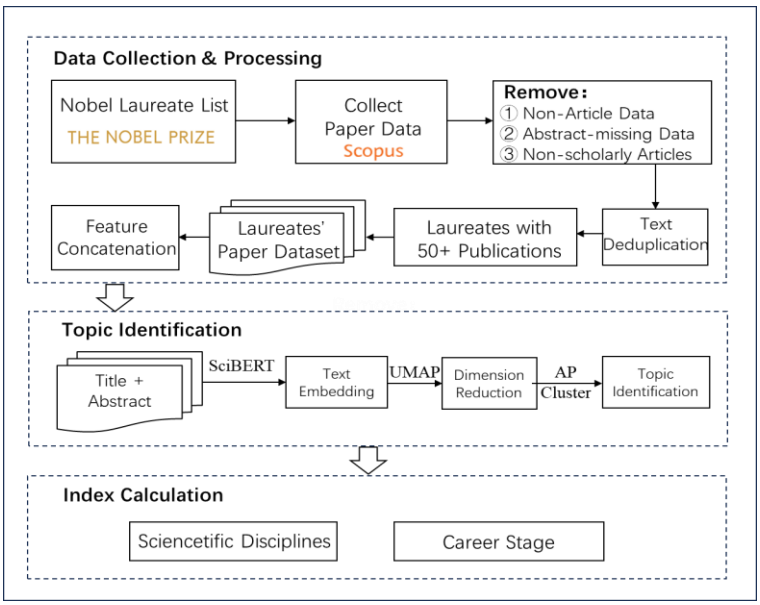


Figure 1. Technology Roadmap.

Data Collection and Cleaning

The Nobel Prize is one of the highest honors in science, awarded to individuals who have made "the greatest benefit to humankind." As elite scientists in different fields, analyzing the characteristics of Nobel laureates can provide valuable insights for the career development of young researchers. Therefore, we select Nobel Laureates from 1901 to 2023 as the research subjects.

Data Collection

First, collect the basic information of Nobel laureates from the Nobel Prize website, including name, award year, etc. Then collect and clean the publication of Nobel laureates. According to the evaluation of author identification effect in WOS, Scopus, AMiner, OpenAlex, and ORCID by Shi Dongbo et al. (2024), Scopus outperforms others in coverage, accuracy, and robustness. Therefore, we use Scopus as the primary data source. Each laureate's personal page is accessed using their Scopus ID, and a dataset of their published papers (retrieved in June 2024) is downloaded, including the title, abstract, keywords, publication date and so on.

Data Cleaning

The collected data is further cleaned as follows: (1) only records classified as "Article" are retained; (2) duplicate entries are removed; (3) records lacking abstract information are excluded; and (4) non-research articles such as Nobel Lectures are removed. Considering that the publication records of some early laureates—especially those awarded before the mid-20th century—may be incomplete, which could affect analysis accuracy, we include only laureates with at least 50 publications. The final dataset comprises 366 Nobel laureates and a total of 82,879 papers, including 123 laureates in Physics (24,116 papers), 123 laureates in Chemistry (32,586 papers), and 120 laureates in Physiology or Medicine (26,177 papers).

Results

Topic Identification Analysis

The distribution of the number of research topics is shown in Figure 2. It can be seen that the number of laureates with 2 research topics in their entire career is the largest, and more than 85% of the laureates have 2 to 5 research topics.

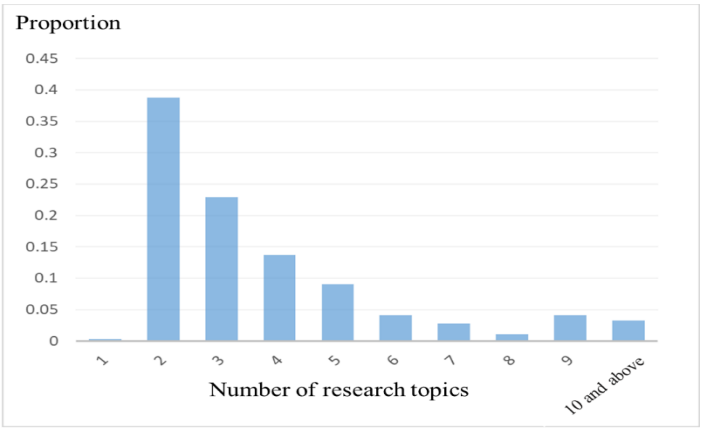


Figure 2. Distribution of the Number of Research Topics of All Laureates.

The average number of research topics of Nobel laureates in different disciplines is shown in Table 1. It can be seen that the average number of research topics of Nobel laureates in Physiology or Medicine in their entire career is the highest, which is 4.2 topics, higher than the average number in chemistry (3.7 topics) and physics (3.4 topics).

Table 1. Average Number of Research Topics of Scientists.

Award Category	*Mean*
Physics	3.4
Chemistry	3.7
Physiology or Medicine	4.2

Index Characteristic Analysis

Each Nobel laureate is calculated separately according to the index calculation method. Then calculate the average to represent the average level of each disciplinary field.

Topic Coverage Index (C)

Figure 3 and Table 2 respectively illustrate the distribution of topic coverage and related statistical indicators for laureates in Physics, Chemistry, and Physiology or Medicine. In terms of average values, the topic coverage among laureates across these disciplines is relatively close. Chemistry laureates show a slightly higher average topic coverage (0.014) compared to laureates in Physics (0.013) and Physiology or Medicine (0.013), suggesting that, overall, Chemistry laureates tend to work on topics with greater differences in content, indicating more diversity and

interdisciplinarity in their research choices. In terms of standard deviation, Physics laureates exhibit the highest variability (0.023), significantly higher than that of laureates in Physiology or Medicine (0.085) and Chemistry (0.047), indicating that Physics laureates show the greatest internal differences in topic coverage within their group.

A closer look at the distribution of topic coverage reveals a right-skewed pattern in all three disciplines, with most values concentrated in the 0–0.02 range. This suggests that most laureates tend to focus on topics with relatively small differences throughout their research careers, concentrating on a limited number of directions. From the perspective of disciplinary differences, Physics laureates exhibit a longer distribution tail, with a maximum value reaching 0.148—substantially higher than the maximums for Chemistry (0.047) and Physiology or Medicine (0.085). This indicates the presence of a small number of Physics laureates whose research spans exceptionally broad areas. One such example is Rainer Weiss, who ranked second in topic coverage and was awarded the 2017 Nobel Prize in Physics for the development of the LIGO detector and the observation of gravitational waves. Over the course of his career, he worked on nine different topics, ranging from particle physics and astrophysics to gravitational wave astronomy, clearly demonstrating strong interdisciplinarity. In Chemistry and Physiology or Medicine, most laureates exhibit lower topic coverage values. However, a small "secondary peak" appears around 0.04, indicating that in these two fields, there is a subset of laureates whose research topics are relatively diverse—though not to the same extent as those in Physics.

Table 2. Descriptive Statistics of Topic Coverage Index (C).

Statistical Indicator	*Physics Laureates*	*Chemistry Laureates*	*Physiology or Medicine Laureates*
Mean	0.013	0.014	0.013
Minimum	0.000	0.000	0.000
Upper Quartile	0.003	0.004	0.004
Median	0.006	0.008	0.007
Lower Quartile	0.014	0.017	0.013
Maximum	0.148	0.047	0.085
Standard Deviation	0.023	0.015	0.016

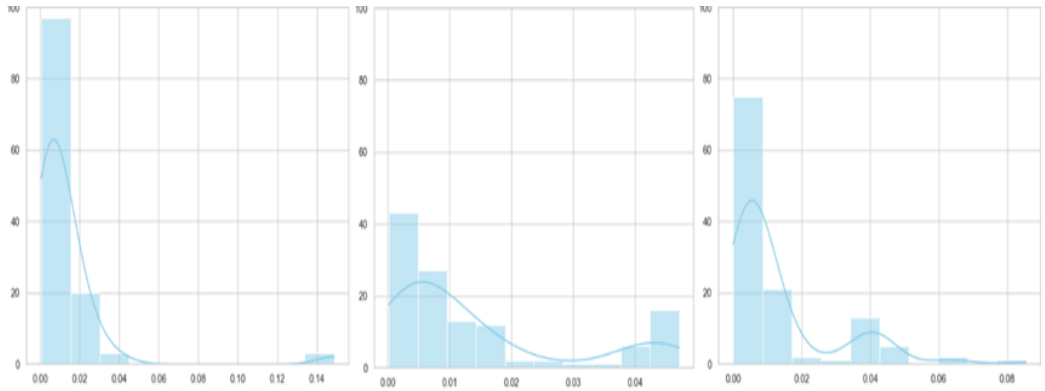


Figure 3. Topic Coverage Index (C) distribution (physics vs. chemistry vs. Physiology or medicine).

Topic Focus Index (F)

As shown in Figure 4 and Table 3, the three disciplinary groups of Nobel laureates exhibit relatively similar patterns in topic concentration. Comparatively, Physics laureates have a slightly higher average topic concentration (0.63) than Chemistry (0.59) and Physiology or Medicine laureates (0.59), indicating that Physics laureates tend to distribute their research efforts more evenly across topics throughout their careers, while Chemistry and Physiology or Medicine laureates tend to focus their research efforts more narrowly. In terms of standard deviation, the topic concentration of Physiology or Medicine laureates shows the highest variation (0.25), slightly higher than that of Physics (0.24) and Chemistry laureates (0.23), suggesting greater within-group differences in how these laureates allocate their research efforts across topics. However, this difference is not particularly significant compared with the other two disciplines.

A closer look at the distribution of topic concentration reveals a general left-skewed trend across all three disciplines, with peaks concentrated in the 0.6–0.8 range. The maximum values of topic concentration in all three disciplines are close to 1 (rounded), indicating that in each field, there are laureates who distribute their research efforts almost equally among multiple topics. For example, Koichi Tanaka, who won the Nobel Prize in Chemistry in 2002, distributed his research efforts nearly evenly across two topics over his career. Topic 1 involved the synthesis and reaction mechanisms of small organic molecules, including cycloaddition and aromatic substitution reactions, with 97 papers. Topic 2 focused on the structure and electronic properties of large conjugated systems, under which he published 95 papers.

Table 3. Descriptive Statistics of Topic Focus Index (F).

Statistical Indicator	*Physics Laureates*	*Chemistry Laureates*	*Physiology or Medicine Laureates*
Mean	0.63	0.59	0.59
Minimum	0.05	0.07	0.00
Upper Quartile	0.51	0.45	0.40
Median	0.66	0.63	0.64
Lower Quartile	0.80	0.76	0.79
Maximum	1.00	1.00	1.00
Standard Deviation	0.24	0.23	0.25

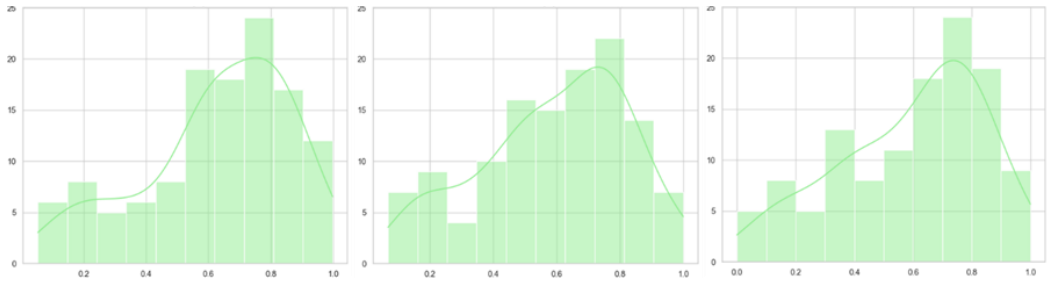


Figure 4. Topic Focus Index (F) distribution (physics vs. chemistry vs. Physiology or medicine).

Topic Duration Index (G)

Figure 5 illustrates the distribution of topic persistence among Nobel laureates in the three disciplines, while Table 4 presents the corresponding descriptive statistics. In terms of average values, Physics laureates have a topic persistence of 0.72, Chemistry laureates 0.71, both higher than that of Physiology or Medicine laureates, which stands at 0.66. This indicates that Nobel laureates generally demonstrate strong research persistence, with Physics and Chemistry laureates showing a particularly prominent tendency to pursue long-term research on a single topic. Regarding standard deviation, Physiology or Medicine and Physics laureates both have a topic persistence standard deviation of 0.17, slightly higher than that of Chemistry laureates (0.14), suggesting that the former two groups exhibit greater internal variability in their research persistence. Chemistry laureates, in contrast, display more consistent persistence overall.

Further analysis of the distribution reveals that the topic persistence of Physics laureates is mostly concentrated above 0.5, with a notable peak around 0.85 and a considerable number of laureates approaching a persistence score of 1. Overall, Physics laureates exhibit high topic persistence, with over 75% of them dedicating

more than half of their career to a single topic—demonstrating a strong commitment to long-term research. For instance, Horst L. Stormer, who won the Nobel Prize in Physics in 1998, had a topic persistence as high as 0.99. He published his first paper in 1976 and had a research career spanning 35 years. During this period, he focused on two main topics: (1) semiconductor materials and electronic transport, including modulation doping and two-dimensional electron gases, and (2) his well-known research on the quantum Hall effect and its physical mechanisms. His research spanned 34 and 35 years on these two topics respectively, exemplifying deep and continuous engagement in specific research areas. Chemistry laureates show a more symmetrical distribution of topic persistence, primarily ranging from 0.5 to 0.9, with a relatively flat peak, suggesting a balanced overall pattern. There are fewer laureates at the extreme low or high ends of the scale. In contrast, the topic persistence of Physiology or Medicine laureates displays a bimodal distribution, with two main peaks: one between 0.5 and 0.6, and another between 0.75 and 0.85. The peak around 0.5 is the highest, indicating that in this field, the largest group of laureates falls into the moderate range of topic persistence—on average, spending about half of their research careers focused on a single topic.

Table 4. Descriptive Statistics of Topic Duration Index (G).

Statistical Indicator	*Physics Laureates*	*Chemistry Laureates*	*Physiology or Medicine Laureates*
Mean	0.72	0.71	0.66
Minimum	0.28	0.37	0.28
Upper Quartile	0.60	0.59	0.52
Median	0.73	0.72	0.65
Lower Quartile	0.86	0.83	0.80
Maximum	0.99	0.99	1.00
Standard Deviation	0.17	0.14	0.17

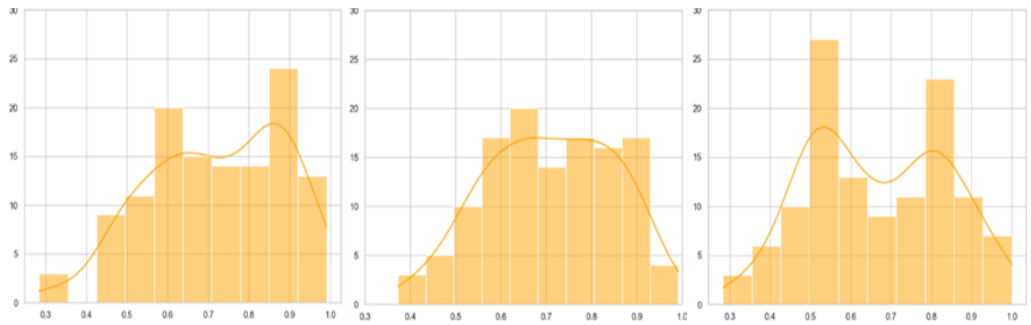


Figure 5. Topic Duration Index (G) distribution (physics vs. chemistry vs. Physiology or medicine).

Topic Shifting Speed

Figure 6 and Table 5 respectively present the distribution and descriptive statistics of topic switching speed among Nobel laureates in Physics, Chemistry, and Physiology or Medicine. In terms of average values, laureates in Physiology or Medicine have the highest topic switching speed at 0.40, followed by Chemistry laureates at 0.37 and Physics laureates at 0.35. This indicates that Physiology or Medicine laureates tend to switch between research topics more frequently throughout their careers, continuously exploring directions different from their current research. In terms of standard deviation, Chemistry laureates exhibit the highest variation in topic switching speed (0.32), suggesting substantial internal differences in how frequently they change research topics. In contrast, Physics laureates show the lowest standard deviation (0.25), indicating a more consistent pattern across the group, with generally lower switching speeds.

Further analysis of the distribution shows that topic switching speed in all three disciplines is significantly right-skewed, with most laureates' switching speeds concentrated between 0.2 and 0.4—particularly in Chemistry. This suggests that most laureates have relatively low switching speeds over the course of their careers, tending to stay focused on their current lines of research. However, the disciplines differ more notably at the extremes. The maximum switching speed among Chemistry laureates reaches as high as 1.62, which is higher than that of Physiology or Medicine laureates (1.33) and Physics laureates (1.26). This implies that a small number of Chemistry laureates change research topics extremely frequently. One example is Roald Hoffmann, a Chemistry Nobel laureate with a switching speed of 1.62. Over the course of his career, he explored 12 different research topics, showing a clear pattern of shifting directions. He began with the development of quantum chemistry and molecular orbital theory, then applied these theoretical methods to the analysis of organic reaction mechanisms. His research later expanded into the structural and reactive properties of inorganic and organometallic compounds, and ultimately extended to the theoretical design of electronic structures in solid-state materials and novel conductive systems.

Table 5. Descriptive Statistics of Topic Shifting Speed (S).

Statistical Indicator	*Physics Laureates*	*Chemistry Laureates*	*Physiology or Medicine Laureates*
Mean	0.35	0.37	0.40
Minimum	0.03	0.04	0.00
Upper Quartile	0.18	0.16	0.18
Median	0.30	0.27	0.33
Lower Quartile	0.44	0.49	0.56
Maximum	1.26	1.62	1.33
Standard Deviation	0.25	0.32	0.29

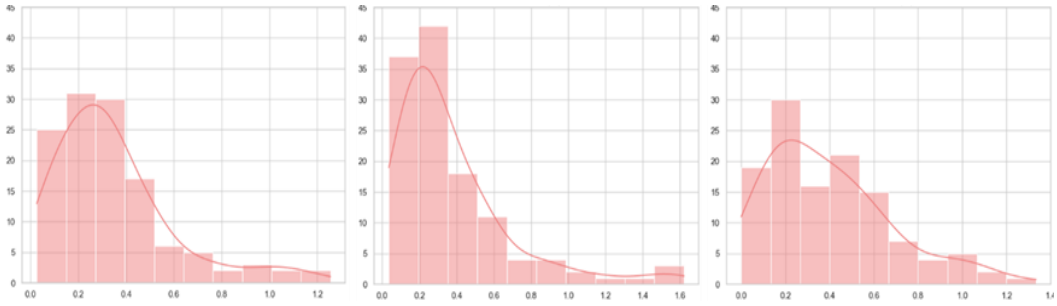


Figure 6. Topic Shifting Speed Index (S) distribution (physics vs. chemistry vs. Physiology or medicine).

Trends in characteristic indicators across different career stages

Figure 7 illustrates the changing trends of characteristic indicators across different career stages of Nobel laureates in Physics, Chemistry, and Physiology or Medicine. It can be seen that, except for the Topic Coverage Index (C), the other three indicators show similar patterns of change.

Figure 7. Trends in characteristic indicators across different career stages.

From the trend in topic coverage, Physics laureates exhibit a distinctive upward

trajectory as their careers progress, indicating that the diversity of their research content increases over time. This suggests enhanced interdisciplinarity and greater variation in their question selection. In contrast, Chemistry and Physiology or Medicine laureates show a declining trend in topic coverage, reflecting a reduction in content diversity and a growing tendency toward thematic convergence.

With respect to topic concentration, all three disciplines demonstrate a gradual downward trend across the career span, with a particularly pronounced decline during the late career stage. This pattern reflects a transition from a relatively dispersed allocation of research efforts in the early career stage to a more focused investment in core topics over time, highlighting an increasing degree of specialization. This also implies that Nobel laureates tend to explore multiple fields in the early stages of their careers, but as their research interests become more defined, they increasingly build upon and deepen their existing research foundations. Notably, Physics laureates consistently maintain a more evenly distributed research effort—not only across their entire career trajectories but also within the early, mid, and late career stages.

In terms of topic persistence, an overall increasing trend is observed across career stages. Physics laureates demonstrate relatively high persistence from the early career stage, which continues to rise steadily as their careers advance. Chemistry and Physiology or Medicine laureates show similar trajectories, with a substantial increase in persistence during the mid-career stage compared to the early stage, followed by a relatively stable level thereafter. This suggests that long-term engagement with specific research topics becomes increasingly prominent as laureates progress through their careers.

Regarding topic switching speed, laureates across all three disciplines tend to exhibit the highest switching rates during the mid-career stage, indicating a greater inclination to explore new directions distinct from their current research. Notably, Physiology or Medicine laureates demonstrate a consistently high switching tendency not only throughout the entire career span but also within the early, mid, and late stages, suggesting a more dynamic and exploratory research pattern within this field.

Conclusion and Discussion, Future work

In this paper, we construct an index system for scientists' question selection and conduct an empirical analysis using Nobel Laureates in natural sciences as examples. This study unveils the multi-dimensional characteristics of scientists' research question selection, offering insights into the research patterns of scientists' question selection.

In conclusion, Nobel laureates share several common characteristics in their research question selection. First, their research topics are typically concentrated in a limited number of directions, with a certain degree of thematic relatedness. Second, they exhibit strong research persistence, often engaging in long-term, in-depth

exploration of a single topic. Third, they tend to strike a balance between focused and dispersed allocation of research efforts. At the same time, there are notable disciplinary differences in question selection characteristics. Compared to other fields, Physics laureates show stronger thematic coherence, more evenly distributed research efforts, longer durations of topic engagement, and lower switching frequency. Their question selection is characterized by sustained and stable advancement across multiple interrelated research areas. Chemistry laureates, by contrast, display greater interdisciplinarity and cross-domain exploration. Although they may switch research directions during their careers, such changes typically revolve around one or a few core themes. Physiology or Medicine laureates exhibit more exploratory question selection patterns, frequently shifting between one or several related core topics, with relatively lower research persistence.

As their careers progress, Physics laureates demonstrate a tendency to both broaden the scope and deepen the focus of their research. This is manifested in increasing interdisciplinarity and diversity in question selection, a gradual concentration of research efforts, and steadily strengthening research persistence. Chemistry and Physiology or Medicine laureates, on the other hand, generally follow a similar trajectory characterized by broad exploration in the early career stage, flexible adjustment in the mid-career stage, and focused deepening in the late career stage. This pattern is reflected in a gradual reduction in the diversity of research content, the emergence of core research themes, and a progressive increase in research persistence, which remains relatively stable in the later stages of their careers. These findings not only shed light on the nuanced evolution of question selection among Nobel laureates across disciplines but also provide valuable insights into the dynamic interplay between research breadth and depth throughout a scientific career.

However, there are still limitations. First, the analysis requires further interpretation and robustness verification. Future work will incorporate qualitative validation through interviews, biographies, and other textual data from scientists at different career stages to enhance the credibility of the results. Second, the focus on Nobel laureates limits the analysis, excluding comparisons with scientists at other levels. Future research will include data from scientists at various levels, such as members of the American Academy of Sciences and other researchers, for a comparative analysis.

References

Aleta, A., Meloni, S., Perra, N., & Moreno, Y. (2019). Explore with caution: mapping the evolution of scientific interest in physics. EPJ Data Science, 8(1). doi:10.1140/epjds/s13688-019-0205-9

Bu, Y., Huang, S., Huang, Y., & Lu, W. (2022). Temporal-Spatial Measurements for Research Topic Evolution of Researchers: Speed, Volume, and Circuitousness. Library and information service, 66(24), 84-91.

Chakraborty, T., Tammana, V., Ganguly, N., & Mukherjee, A. (2015). Understanding and modeling diverse scientific careers of researchers. Journal of Informetrics, 9(1), 69-78.

doi:10.1016/j.joi.2014.11.008
Chen, L., Guo, S., Teng, G., & Tuo, R. (2019). Research Topic Focus and Transfer of Scientific Personnel. Digital Library Forum(12), 9-17.
Chen, X., Pan, Y., Ma, Z., Zhang, G., Zhang, B., & Ren, Q. (2023). The Influence of Fund Support on the Research Direction of Outstanding Young Scholars. Journal of the China Society for Scientific and Technical Information, 42(12), 1438-1447.
Ding, J., & Chen, Y. (2023). Concentration, continuation and Extension: Research Pattern of Nobel Laureates Based on the Topic: An Empirical Analysis of Physics. Library Journal, 42(08), 100-109.
Ding, J., Chen, Y., & Liu, C. (2023). Exploring the research features of Nobel laureates in Physics based on the semantic similarity measurement. Scientometrics, 128(9), 5247-5275. doi:10.1007/s11192-023-04786-3
Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and Innovation in Scientists' Research Strategies. American Sociological Review, 80(5), 875-908. doi:10.1177/0003122415601618
Huang, S., Huang, Y., Bu, Y., Luo, Z., & Lu, W. (2023). Disclosing the interactive mechanism behind scientists' topic selection behavior from the perspective of the productivity and the impact. Journal of Informetrics, 17(2). doi:10.1016/j.joi.2023.101409
Jia, T., Wang, D., & Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. Nature Human Behaviour, 1(4). doi:10.1038/s41562-017-0078
Liu, F., & Xia, H. (2017). Discovering Interest Transition Patterns of Technological Inventors. Technology Intelligence Engineering, 3(02), 33-40.
Liu, L., Jones, B. F., Uzzi, B., & Wang, D. S. (2023). Data, measurement and empirical methods in the science of science. Nature Human Behaviour, 7(7), 1046-1058. doi:10.1038/s41562-023-
Liu, M., Shi, J., Yang, S., & Bu, Y. (2024). Speed of Research Topic Evolution and Scientific Performance: Evidence from Computer Science. Library and information service, 68(06), 72-82.
Meng, Z. (1994). General psychology. Beijing: Peking University Press.
Shi, D., Deng, H., Yang, Z., Liu, N., Liu, Y., & Mao, Y. (2024). A Study on the Applicability of Author Identification Numbers in Scientific and Technical Paper Databases. ChinaXiv. doi:doi:10.12074/202406.00022V2
Taylor, J. A., Larraondo, P., & de Supinski, B. R. (2022). Data-driven global weather predictions at high resolutions. INTERNATIONAL JOURNAL OF HIGH PERFORMANCE COMPUTING APPLICATIONS, 36(2), 130-140. doi:10.1177/10943420211039818
Van Houten, J., Van Vuren, H., Le Pairs, C., & Dijkhuis, G. (1983). Migration of physicists to other academic disciplines: situation in the Netherlands. Scientometrics, 5(4), 257-267.
Yu, X., Szymanski, B. K., & Jia, T. (2021). Become a better you: Correlation between the change of research direction and the change of scientific performance. Journal of Informetrics, 15(3). doi:10.1016/j.joi.2021.101193
Zeng, A., Shen, Z., Zhou, J., Fan, Y., Di, Z., Wang, Y., . . . Havlin, S. (2019). Increasing trend of scientists to switch between topics. Nat Commun, 10(1), 3439. doi:10.1038/s41467-019-11401-8

https://doi.org/10.51408/issi2025_088

Exploring Novelty Differences between Industry and Academia: A Knowledge Entity-centric Perspective

Hongye Zhao[1], Yi Zhao[2], Chengzhi Zhang[3]

[1]*zhaohongye_phd@njust.edu.cn*, [2]*yizhao93@njust.edu.cn*, [3]*zhangcz@njust.edu.cn*
Department of Information Management, Nanjing University of Science and Technology, Nanjing (China)

Abstract

Novel ideas drive innovation, and both academia and industry possess distinct strengths in advancing technological progress. The industrial sector, on the one hand, seeks to privatize knowledge to maintain appropriability, while on the other hand, it actively promotes open-sourcing of models and platform sharing. This paradox raises the question of whether industrial disclosures are less novel compared to those from academia. Some studies argue that academia tends to generate more novel ideas, while others suggest that industry researchers are more likely to drive new breakthroughs. Previous studies have been limited by data sources and inconsistent measures of novelty. To address these gaps, this study establishes a unified framework for calculating the novelty of papers and patent data in the field of Natural Language Processing (NLP), focusing on fine-grained knowledge entities. Additionally, a regression model is constructed to analyse the relationship between the type of institution and the novelty of their publications. The results show that academia demonstrates higher novelty in both patent and paper outputs. Notably, academic involvement significantly enhances the novelty of industrial patents. Furthermore, this study examines how team size impacts novelty in patents and papers, providing strategic recommendations for forming research teams. We release our data and associated codes at https://github.com/tinierZhao/entity_novelty.

Introduction

Academic research focuses on theoretical inquiry and the advancement of fundamental lence, aiming to expand human knowledge and drive disciplinary progress (Sauermann & Stephan, 2010). In contrast, the industrial sector emphasizes core competitiveness (Geisler, 1995), prioritizing economic returns and often safeguarding intellectual appropriability by restricting the disclosure of research outcomes (Arundel, 2001; Chirico et al., 2018).

Following this logic, the industry would typically choose to limit the disclosure of novel research outcomes to safeguard its competitive advantage. However, this traditional notion is being challenged in the field of artificial intelligence (AI), as the industry demonstrates a noticeably more open attitude. For example, leading tech companies have released cutting-edge technologies in algorithms and models, such as the BERT model (Devlin et al., 2019) and various other open-source large language models. Additionally, they have significantly lowered the barriers to adopt artificial intelligence technologies by offering application programming interfaces (APIs) and detailed technical documentation. This enables users to easily integrate these models into their own projects and supports further development and customization. Moreover, the industrial sector's active participation in most active and popular AI conferences has spurred numerous disruptive innovations (Liang et al., 2024). While this openness may partially diminish the appropriability of

knowledge, it offers substantial benefits. On the one hand, the public release of frontier research attracts a broader developer community, reducing long-term maintenance and development costs while generating economic returns through technology services (Homscheid et al., 2015). On the other hand, collaborations with prominent enterprises and academic institutions allow the industrial sector to access external knowledge, thereby maintaining its technological leadership and fostering product iteration and optimization through knowledge spillovers (Jiang et al., 2024), which in turn serves to broaden its market share (Hu et al., 2023; Tao et al., 2022). In this context, it remains uncertain whether the research outcomes from industry exhibit lower novelty compared to those from academia.

Evaluating the novelty of scientific and technical literature presents inherent challenges. Publications that introduce revolutionary technologies and lay the foundation for subsequent studies are rare (Arts et al., 2019; Arts et al., 2021). These works often go underappreciated initially, as they challenge existing conventions and may encounter resistance during the review process (Riera & Rodríguez, 2022). In contrast, studies that align with established theories are more likely to gain peer trust (Liang et al., 2022), putting highly novel research at a disadvantage in peer review (Koppman & Leahey, 2019; Wang et al., 2017). Even after publication, such research often faces delays in gaining recognition (Wang et al., 2017). Meanwhile, the growing volume of scientific and technical literature across disciplines has significantly increased the workload for reviewers (Shibayama et al., 2021).

In this context, the novelty of industrial disclosures compared to academic ones remains a topic of ongoing debate. As a key branch of artificial intelligence, NLP continues to experience rapid growth, with significant breakthroughs emerging from both academia and industry, despite a general slowdown in innovation across many fields (Park et al., 2023). Some scholars argue that academia contributes more novel ideas, while industry tends to adopt and refine academic advancements (Bikard & Marx, 2019). Subsequent studies further confirm academia's leadership in NLP innovation (Chen et al., 2024; Liang et al., 2024). However, Dwivedi et al. (2019) suggest that industry researchers are more likely to drive new AI technologies. The rise of pre-trained models such as Transformer (Vaswani et al., 2017) and GPT (Radford et al., 2018), along with the rapid development of large-scale language models like ChatGPT, Ahmed et al. (2023) highlights industry's dominance in computational resources, data, and talent.

To date, studies on the differences in the novelty of publications between academia and industry in NLP have primarily focused on papers. The limitation is not due to the availability of data. Instead, it occurs because the approaches for evaluating novelty vary considerably between patents and scientific papers. For scientific papers, novelty is typically measured through journal citation pair analysis. However, patents primarily cite other patents rather than academic papers (Ba et al., 2024), and they do not correspond to journal types. Therefore, the novelty of patents cannot be directly measured using citation journal pairs. Moreover, the classification codes commonly used in patents cannot be aligned with those used in scientific papers. As a result, previous studies have not fully incorporated patent data, leaving a gap in understanding the specific relationship between institutional types and the novelty of scientific and technical literature.

This study addresses the gap by using a unified novelty evaluation framework that leverages fine-grained knowledge entities to assess the novelty of publications across academia, industry, and their collaborations in NLP. We calculate the semantic distances between fine-grained knowledge entities and assess the difficulty of different entity combinations. Unlike previous studies, this research selects specific entity types based on the characteristics of the NLP field, reducing interference from certain types and enhancing the reliability of novelty assessments. While focused on NLP, the methodology is applicable to other domains, particularly those involving scientific papers and patents outcomes. It offers a general analytical framework for comparing novelty across academia and industry and evaluating the effectiveness cross-sector collaboration.

Specifically, we address the following two research questions:

RQ1: How to unify the novelty calculation method based on fine-grained knowledge entities for both papers and patents?

RQ2: Is there a difference in the novelty of scientific and technical literature between industry and academia?

The contributions of this paper are as follows:

First, we extend the entity-based novelty measurement method to the patent domain. By transferring the entity recognition model from papers to patents, we apply the same novelty measurement to both, enabling a unified assessment and supporting future data source expansion.

Second, our analysis confirms that academic outputs in the NLP field exhibit higher novelty. Additionally, our findings indicate that patents generated through collaboration between industry and academia exhibit a significant increase in novelty, highlighting the potential impact of cross-sector collaboration on innovation.

The code and data used in this study are open-sourced on GitHub and can be accessed via the following website: https://github.com/tinierZhao/entity_novelty

Related work

For the research questions proposed in this paper, we conducted a review of the scientific and technical literature on novelty measures, as well as the factors influencing novelty.

Novelty measures in the scientific and technical literature

The measurement of novelty not only helps to identify valuable innovations in advance, but also provides key insights for technological transfer and innovation. Currently, novelty is primarily measured through combinations, as Nelson and Winter (1982) argued, "the creation of novelty mainly involves the recombination of existing conceptual and physical materials." Traditional methods for measuring novelty include the use of journal pairs and classification code pairs to assess the novelty of literature. With the availability of large-scale data and the advancement of machine learning and natural language processing technologies, novelty measurement methods have been continuously innovated. The combination of other types of knowledge elements has gradually become an important approach for

assessing novelty. Additionally, some studies have explored new avenues by treating novelty as a binary classification task, using classification or outlier detection methods to distinguish between novel and non-novel literature.

From a combination-based view, early methods primarily focused on citation references and classification codes. Uzzi et al. (2013) compared the observed and Monte Carlo-simulated frequencies of journal pairs to calculate z-score for each pair, using the lowest 10th percentile z score to indicate a paper's novelty and the median z score to indicate its conventionality. Lee et al. (2015) improved Uzzi's method in terms of computational difficulty by adopting a multi-year time window, which reduced the previous single-year window and calculated the commonness of citation pairs. Wang et al. (2017) measured novelty through the first-time combination of different citation journal pairs in a paper. Specifically, they constructed a co-citation matrix for the journals and used cosine similarity between the vectors of each journal to assess the difficulty of combining the journal pairs. However, while these methods are easy to understand and explain, they also face limitations such as self-citation and biased citing (MacRoberts & MacRoberts, 1996; Jeon et al., 2023; Anne, 2023). Additionally, as the number of papers analysed increases, costs and computational efficiency escalate sharply.

Regarding patent novelty measurement, early traditional methods focused on patent classification codes and backward citations (Ahuja & Lampert, 2001; Lee & Lee, 2019). However, citations merely describe existing technologies and fail to reflect the technology of the patent itself, often presenting incomplete and biased representations (Kuhn et al., 2020; Arts et al., 2021). Measuring technological novelty through patent IPC codes (Fleming, 2001) is overly broad and tends to capture interdisciplinarity rather than technological uncertainty.

With the continuous development of NLP technologies, tasks such as scientific terminology extraction (entities, keywords) and semantic embedding have matured, making the measurement of novelty based on scientific text content a more reasonable and effective approach. Liu et al. (2022) used the BioBERT model to calculate the semantics of biological entities, determining entity pair novelty based on semantic similarity. The novelty score for each paper is calculated as the proportion of novel entity pairs to the total possible entity pairs. Similarly, Chen et al. (2024) applied an entity similarity-based approach using S to evaluate the novelty of conference papers in the field of natural language processing. Luo et al. (2022) employed BERT word embeddings to measure novelty by assessing the novelty of research questions, methods, and their combinations. Arts et al. (2021) extracted keywords from patent titles and abstracts, calculating “new_ngram” and corresponding “new_ngram_reuse” to measure patent novelty. Wei et al. (2024) used the BERT model to extract innovative sentences from patent claims and distilled them into knowledge element triples, measuring novelty scores for the triples by projecting entities and relations into a common space, using a combination of word2vec and HGT.

Table 1. Related works of novelty measurement.

Author	*Domain and data*	*Method*
Uzzi et al. (2013)	17.9 million papers spanning all scientific fields	Monte Carlo + Journal pairs combinations
Wang et al. (2017)	785,324 Articles in 251 subject	Co-citation matrix + Journal pairs combinations
Liu et al. (2022)	98,981 coronavirus papers	BioBERT + Knowledge entities combinations
Chen et al. (2024)	14,812 ACL Anthology papers	SciBERT + Knowledge entities combinations
Wei et al. (2024)	1343 agricultural robots patents	BERT + Knowledge triples combinations
Luo et al. (2022)	204,224 papers in ACM database	BERT + Questions-Methods combinations
Arts et al. (2021)	1,302,956 patents spanning all fields	SnowBall + New_ngram combinations
Jeon et al. (2022)	1,877 medical image patents	Doc2Vec + Outlier detection binary classification
Jeon et al. (2023)	15,653 biomedical papers	FastText + Outlier detection binary classification
Zanella et al. (2021)	13,393 blockchain-related patents	Word2Vec + Outlier detection binary classification
Jang et al. (2023)	25,183 pairwise vehicle communication networks patents	RoBERTa + Explainable AI binary classification

From the perspective of binary classification. Jang et al. (2023) treated patent novelty as a classification task, using RoBERTa for semantic embedding of patent claims to develop a self-explainable novelty classification model. Jeon et al. (2022) embedded patent claims and used the local outlier factor (LOF) algorithm to calculate patent novelty. Their study showed that, although ELMo and BERT provide high-quality patent embedding vectors, they are less suitable for modeling the technological features of patents, particularly in single technical domains, compared to Doc2Vec. Jeon et al. (2023) trained a fastText model using paper titles

in the biomedical field and applied the LOF algorithm to measure the novelty score of each paper. Zanella et al. (2021) combined cosine similarity and density-based anomaly detection to improve the identification of outliers within patent clusters. A detailed summary of the above works, including their data, methods, is provided in Table 1.

From the above-mentioned studies, the methods for measuring novelty have evolved from the early approaches relying on citation and classification codes to those based on text content analysis. Moreover, no unified framework yet exists for calculating the novelty of patents compared to scientific papers. In the following chapters, we will provide a detailed explanation of how to uniformly extract fine-grained knowledge entities from patents and papers, and how to calculate the novelty based on fine-grained knowledge entities.

Factors influencing the novelty of scientific and technical literature

Previous studies have explored the relationship between novelty from various perspectives, including institutional nature, team size, and author attributes within teams.

Regarding team size, existing research presents inconsistent findings. Uzzi et al. (2013) found that research teams are more likely to introduce novel combinations within familiar knowledge domains compared to single-author papers. Lee et al. (2015) identified an inverted U-shaped relationship between team size and novelty, with this effect largely driven by the interplay between team size and knowledge diversity. Wang et al. (2019) suggested that smaller teams are more likely to disrupt science and technology with new ideas, while larger teams tend to focus on existing ones. Shin et al. (2022), using Web of Science data, found that scientific collaboration negatively affects novelty, as collaborative research tends to remain within established fields. However, Wu et al. (2024) argued that collaboration fosters trust and problem-solving abilities, and that knowledge diversity enhances knowledge transfer and promotes the impact of science on technology. Conversely, some studies indicate that excessive team heterogeneity may reduce trust, hinder knowledge sharing, and obstruct innovation (Chen et al., 2015).

At the institutional level, academia tends to lead industry in terms of novelty at the paper level, generating more exploratory ideas, while industry is more likely to produce high-impact papers (Liang et al., 2024). Chen et al. (2024) measured the novelty in the NLP field, finding that academia and collaborative institutions tend to be more novel than industry, based on fine-grained combinations of knowledge entities. Other studies suggest that papers involving companies have a higher impact, and collaborations between industry and academia exhibit greater novelty (Jee & Sohn, 2023).

At the author attribute level within teams, teams with diversified expertise tend to produce more original work and have a long-term advantage in terms of impact (Zheng, Li, & Wang, 2022). Mori and Sakaguchi (2018) examined how differentiated knowledge among inventors enhances patent novelty using Japanese patents. Gender diversity within teams has also become a favored topic in recent years. Teams with gender diversity produce papers with higher novelty and greater

impact compared to single-gender teams (Yang et al., 2022). Liu et al. (2024) explored the relationship between novelty and gender heterogeneity in doctoral theses, finding that female authors had lower average novelty scores than male authors, and male advisors were more likely to supervise students who produced theses with higher novelty. Notably, this gender difference was more pronounced in lower-prestige universities. Similarly, Chan and Torgler (2020) found that among elite scientists, female scientists tend to receive more citations than their male counterparts.

In this study, we explore the performance of different institutional types in terms of novelty in patents and papers, with a particular focus on comparing the relationship between novelty and team size, to uncover both consistencies and differences.

Methodology

This study aims to quantify the impact of different team compositions on the novelty of scientific and technical literature. The research framework in Figure 1 outlines three key steps:

First, dataset construction. We constructed an original dataset that includes scientific and technical literature in the NLP field, comprising papers and patents published between 2000 and 2022, and extracted author information and their affiliated institutions for each document.

Second, novelty assessment of scientific and technical literature. Fine-grained knowledge entities were extracted from both scientific papers and patents, with the knowledge from scientific papers being transferred to patents. To achieve this, we first employed an entity recognition model trained on scientific papers to perform preliminary entity extraction from patent texts. Subsequently, we conducted manual reviews and added annotations for tool-specific terms (such as software platforms) that are unique to patent texts. This iterative process continued until the model's performance converged. The difficulty of their combinations was measured based on the semantic distances between these entities (Liu et al., 2022; Chen et al., 2024). This approach was then used to assess the novelty of each document. Lastly, regression analysis. A regression model was employed to conduct statistical tests on the novelty of scientific and technical literature from different institutions (Chen et al., 2024). Additionally, we treated the top 10% of papers and patents each year as high-novelty documents and performed a robustness check of our results using binary logistic regression (Jeon et al. 2022).

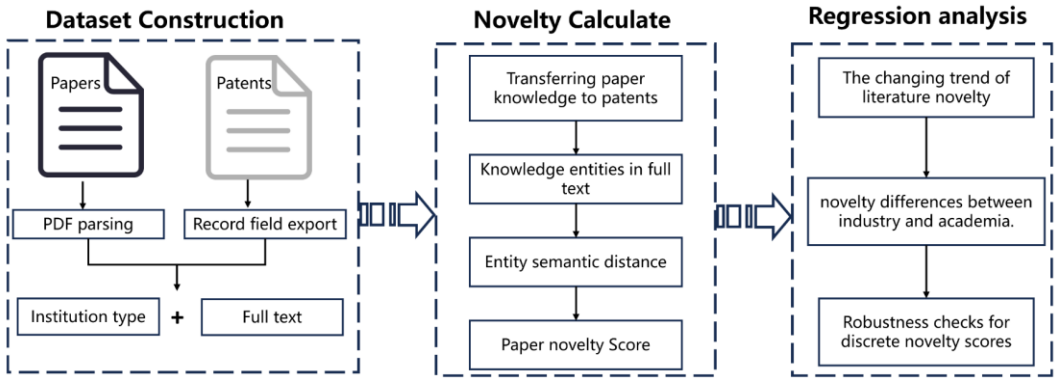


Figure 1. Framework of this study.

Data collection

The paper data was collected from the ACL Anthology[1] website. We selected three representative conferences for our study: ACL (Annual Meeting of the Association for Computational Linguistics), EMNLP (Conference on Empirical Methods in Natural Language Processing), and NAACL (North American Chapter of the Association for Computational Linguistics). A total of 17,783 full-text papers from 2000 to 2022, were collected.

The patent data was collected from the United States Patent and Trademark Office (USPTO) through the patsnap[2] system. We conducted a search for patents within the time frame of 2000 to 2022, using the following query: CPC_GROUP:(G06F40 [3]) AND APD: [20000101 TO 20221231] AND COUNTRY: ("US"). We focused on invention patents and filtered out those with legal statuses such as withdrawal, rejection, abandonment, application termination, or complete invalidation. Additionally, patents with the same priority were consolidated into families. Ultimately, a total of 25,305 patents were obtained.

Identification of the publishing institution type of the literature

By parsing the full-text PDFs and integrating data from the GRID and OpenAlex databases, we identified the authors and their institutions for 17,783 papers. For institutions not found through the search, we manually supplemented the data. Following Chen et al. (2024) and Xu et al. (2022), we categorized the institutions. In cases of multiple affiliations, we adopted the method of Hottenrott et al. (2021), considering the first-listed institution as the author's primary affiliation.

The patent data processing begins with extracting standardized applicant information from databases, where all non-personal names are presented in either Chinese or English. An edit distance algorithm, combined with a local dictionary, is then applied to normalize institutional names. Based on lexical features, two sets of keywords were defined: one for academic institutions and one for industrial organizations, covering both English and Chinese terms. The algorithm classifies

[1] https://aclanthology.org/
[2] https://www.patsnap.com/
[3] CPC: G06F40, Handling natural language data

institutions containing education-related terms (e.g., "edu," "univer") as academic, and those with company-related terms (e.g., "inc," "ltd," "lp") as industrial. This method ensures efficiency and accuracy, as the database provides standardized applicant fields. For unrecognized institutions, spacy[4] named entity recognition is used to determine whether the applicant is individual. For individual applicants appearing more than twice, we validate with ChatGPT to check for missed categorizations. Finally, the results are manually reviewed to correct and supplement the algorithm's output.

Specially, a paper is classified as "Academia" if all its authors are affiliated with academic institutions (such as universities or research institutes), as "Industry" if all authors are affiliated with industry institutions (such as companies or corporations), and as "Cooperation" if it involves authors from both academia and industry.

The specific institutional distribution for papers and patents is shown in Table 2.

Table 2. The institutional distribution of scientific and technical literature.

Institution Types	*Count*	*Ratio(%)*	*Count*	*Ratio(%)*
	Paper		Patent	
Academia	11,670	65.62	468	1.85
Industry	1,679	9.44	21732	85.97
Cooperation	4,315	24.26	69	0.27
Individual	0	0	2932	11.59
Other	119	0.67	104	0.41

Extraction of fine-grained knowledge entities from papers and patents

We adopt a combinatory perspective to assess the novelty of scientific and technical literature. Specifically, we analyze this based on the characteristics of the NLP field. NLP is a research domain centered around methods and data, with most studies typically involving the following key elements: 1) dataset construction or selection, often involving text resources such as corpora and dictionaries, which serve as the foundation for model training and validation; 2) method selection and application, which defines the strategies and steps for solving problems; 3) the choice of evaluation metrics, used to measure model performance and task quality; 4) the use of tools, including programming languages, software, and open-source tools required for implementing and testing NLP methods (Zhang et al., 2024; Pramanick et al., 2024). Based on this framework, we extract fine-grained knowledge entities from each patent and paper, covering the categories of Method, Tool, Metric, and Dataset.

In the fine-grained knowledge entity recognition task, we used the pre-trained SciBERT model. Due to differences in writing style and text structure between patents and papers, we trained separate entity recognition models for each type of document. Specifically, for papers, we adopted the framework proposed by Zhang

[4] https://pypi.org/project/spacy/

et al. (2024). For patents, we initially applied a pre-trained model to annotate the patent sections, followed by re-annotation of the extracted entities according to the labelling rules. Additionally, for unique entities in patent texts, such as Storage medium, we performed extra annotation. After several rounds of iteration and adjustments, we obtained the patent entity recognition model (SciBERT + CRF), which achieved the following performance: Precision of 78.83%, Recall of 82.51%, and F1 score of 80.63%. Given that extracting entities only from titles and abstracts would miss many, we performed full-text extraction for both patents and papers. Paper data were extracted from PDFs, and the patent database was also exported in full text. For entity normalization, we used edit distance and semantic distance to cluster entities. Ultimately, we identified 22,871 entities in the papers and 9,523 entities in the patents.

Table 3. Top 5 entities in four types extracted from papers and patents.

Type	*Paper*		*Patent*	
	Entity	*Frequency*	*Entity*	*Frequency*
Method	BERT	4159	Neural network	3021
	Transformer	3844	Machine learning	1608
	N-gram	3733	N-gram	1365
	LSTM	3607	Language models	1160
	Attention Mechanism	3425	Deep learning	960
Tool	Pytorch	730	Computer system	11646
	MOSES	647	Storage medium	10413
	GIZA++	581	User interface	9323
	Python	430	Computer program	8738
	NLTK	333	Operating system	7636
Dataset	Wikipedia	3534	Emoji	306
	WordNet	2661	Email	122
	Twitter	1324	Soial meida	86
	Wall Street Journal	1005	World wide web	67
	Amazon Mechanical Turk	883	Twitter	43
Metric	Accuracy	10784	Accuracy	5278
	F_1	7802	Confidence	2500
	Precision	6024	Efficiency	2195
	Recall	5551	Relevance	1612
	Confidence	3832	Error	1453

Table 3 presents the top 5 entities in each category for both patents and papers. Due to the fact that patents are rarely evaluated on public datasets, the proportion of Dataset entities in patents is quite low, and as a result, the recognition performance for these entities is somewhat weaker. Additionally, a distinctive feature of patent

terminology is its level of abstraction, particularly evident in the claims section. Unlike general discourse, which relies on precise wording to accurately convey content and avoid vague or overly broad terms, patent claims intentionally use generalized vocabulary (Codina-Filbà et al., 2016). This strategy enables companies to broaden the scope of their intellectual property protection, ensuring more extensive exclusivity over their innovations (Arinas, 2012; Ashtor, 2021). Furthermore, descriptions of Tool entities in patents tend to be more generalized, reflecting this situation.

Measurement of scientific and technical literature novelty

We explore the novelty of entity combinations through an analysis of the fine-grained knowledge entities extracted from scientific and technical literature. We draw on the work of Liu et al. (2024) in the field of scientific novelty assessment for biomedical papers. They treated biological entities as core elements of the research method and used the pre-trained Bio-BERT model to quantify the semantic distance between these entities to measure novelty. We applied this approach to evaluate the novelty of papers and patents in NLP, using pre-trained SciBERT to calculate the semantic similarity of entities for novelty measurement. Specifically, we extracted embeddings for each entity word from the "last_hidden_state", removing [CLS] and [SEP] tokens. If an entity tokenizer contains multiple subwords, we averaged their embeddings. Cosine similarity was then used to calculate their semantic similarity. We labeled the top 10% of entities with the highest semantic distance as high-novelty entities. Finally, we analysed the frequency of these high-novelty entities in the text and measured the novelty of each paper based on their proportion in all entity combinations.

Furthermore, in domain-specific entity analysis, the ubiquity of certain entity types can cause inconsistencies between semantic distance and the actual difficulty of combining entities. For example, entities in the Metric category (such as accuracy, precision, recall, F_1 score, etc.) are often highly generic and strongly associated with most methods, but their semantic distance may not accurately reflect the actual situation. Due to their widespread use, these entities contribute little to novelty measurement and may even introduce noise. Therefore, we excluded entities of the Metric category from our analysis. For an entity pair (e_i, e_j), the distance between the two is denoted as D, and $cosine(e_i, e_j)$ represents the semantic similarity between the entities. As shown in Equation (1):

$$D(e_i, e_j) = 1 - cosine(e_i, e_j) \quad (1)$$

Regression model for novelty comparison

To investigate the differences in novelty across various institutions, this study employs regression analysis to quantify and compare the novelty demonstrated in the scientific and technical literature produced by different institutions. The following sections provide a detailed description of the process of variable selection and the construction of the regression model.

Dependent variables: In the setting of independent variables, we first use the continuous novelty indicator (Novelty Score) calculated in the previous section for analysis. This indicator measures the proportion of novelty entity combinations in each paper or patent, with a score range from 0 to 1, where a higher score indicates greater novelty. Meanwhile, considering the uncertainty of novelty outcomes, we categorize the top 10% of papers and patents ranked by score each year as high novelty and construct a binary classification variable (Novelty Score 10%) for robustness checks.

Independent variables: This study defines the independent variables as the type of institution. After excluding institutions categorized as "other" and "individual", the remaining institutions are classified into three categories: academia, cooperation, and industry. Specifically, two binary variables—Academia and Cooperation, are defined. The Academia variable is set to 1 if the literature belongs to an academic institution, and the Cooperation variable is set to 1 for literature from cooperative institutions, with both variables set to 0 for literature from industry.

Control variables: In addition, the study considers several control variables to account for team characteristics. Specifically, it first considers the number of institutions (Institutions num), followed by the number of authors for papers and inventors for patents (Au/In num), in order to isolate the pure effect of institution type on the novelty of papers and patents. For patents, we also include the size of the patent family (Family size), which is commonly associated with welfare value and technological impact (Kabore & Park, 2019; Wu et al., 2015). Furthermore, the number of IPC classification codes at the subgroup level (IPC num) is controlled to account for the diversity of the patent's knowledge components (Sun et al., 2022). Finally, we include year as a dummy variable, using the publication year for papers and the application year for patents, to control for potential year-related differences that could affect the results. The summary statistics of the variables and the correlation coefficients between the variables are presented in Table 4. and Figure 2, respectively.

We found a strong correlation between the continuous and discrete forms of the dependent variable (novelty), while the correlations between the independent and dependent variables were weak. We then calculated the variance inflation factors (VIFs) for all explanatory variables to assess multicollinearity. The VIF for papers was 2.79 and for patents was 1.07, both below the threshold of 5 (Marcoulides & Raykov, 2019). These results indicate that multicollinearity has minimal impact on our model, ensuring the reliability of the estimates.

Table 4. Summary statistics of variables for regression analysis (N = 22,269 patents, N = 17,664 papers).

Variable	*Mean*	*Std. Dev.*	*Min*	*Max*	*Mean*	*Std. Dev.*	*Min*	*Max*
	Paper				Patent			
Novelty Score	0.10	0.07	0	0.51	0.11	0.09	0	0.75
Novelty Score 10%	0.10	0.30	0	1	0.10	0.30	0	1
IPC num	-	-	-	-	1.94	1.00	1	10
Family size	-	-	-	-	2.10	1.94	1	82
Au/In num	3.76	2.22	1	77	3.28	2.14	1	26
Institutions num	1.80	1.22	1	44	1.05	0.43	1	15
Academia	0.66	0.47	0	1	0.02	0.14	0	1
Cooperation	0.24	0.43	0	1	0.00	0.06	0	1

Note: The papers do not include IPC numbers or Family size, which are represented as '-'.

Regression analyses: Multivariable regression was conducted to examine how different types of institutions influence the novelty scores of the literature. As shown in Equation (2):

$$Novel_i = \alpha + \beta_1 Academia_i + \beta_2 Cooperation_i + Controls + Y_i + \varepsilon \quad (2)$$

Where $Novel_i$ represents the novelty score of each literature i. The independent variables $Academia_i$ and $Cooperation_i$ indicating whether the literature is from an academic or cooperative institution, respectively. The variable Controls includes a set of control variables, Y_i denotes the publication year, and ε represents the error term in the model.

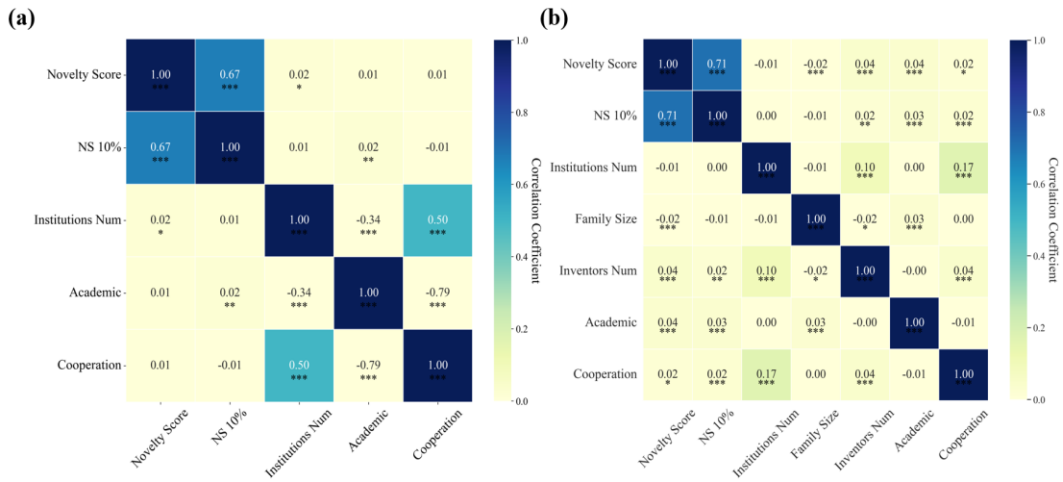


Figure 2. Pearson's rank correlation coefficient matrix (a) Correlation between variables in papers (b) Correlation between variables in patents.

Results

This study analyses papers published between 2000 and 2022 in three major NLP conferences and patents filed with the USPTO, focusing on the novelty differences across three types of publishing institutions: academia, industry, and collaboration. Our research not only compares the performance of different institution types in terms of novelty in literature, but also investigates the relationship between team size and novelty. The aim is to reveal how team size influences innovation across different types of scientific and technical literature.

Trends in publication volume of papers and patents

The field of NLP has experienced rapid growth, with a steady annual increase in patents and papers since 2000. The slight decrease in patent numbers in 2022 compared to 2021 is due to the America Invents Act (AIA), Section 35 U.S.C. § 122(b), which requires patents to be published 18 months after the earliest filing date, unless the applicant requests early publication. As of the retrieval date, some 2022 patents had not yet been published, which is common.

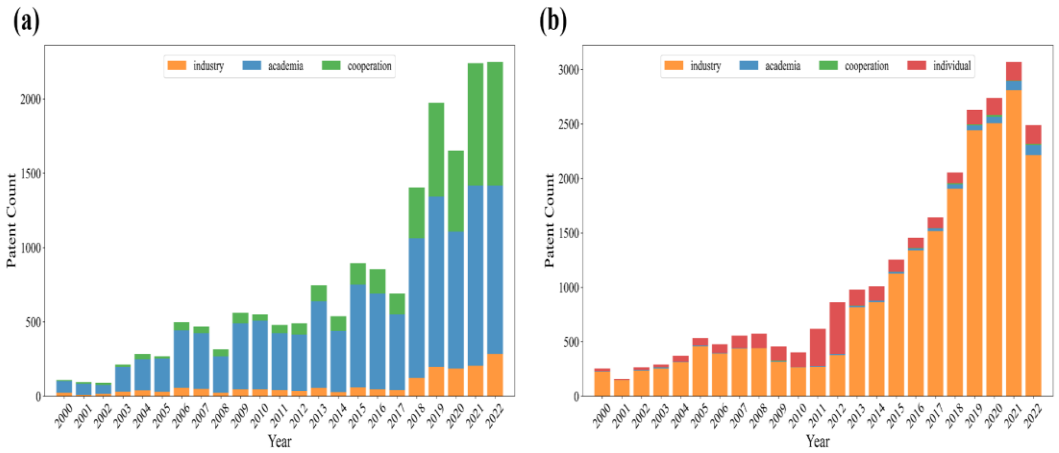


Figure 3. Annual publication volume of papers and patents. (a) Annual publication volume of papers (b) Annual publication volume of patents.

In addition, the distribution of patent numbers across institutions is more uneven compared to papers, with specific proportions detailed in the previous section on institutional distribution. Despite the concentration of the world's top higher education resources in the United States and the majority of government research funding directed towards universities, university-originated patents account for less than 4% of the total national patents, with corporate patents dominating the majority, followed by individual applications[5]. This phenomenon highlights the dominant role of industry in NLP patent filings. The annual publication volume of papers and patents is shown in Figure 3.

[5] https://ncses.nsf.gov/pubs/nsb20204/invention-u-s-and-comparative-global-trends

Trends in novelty changes of literature measured under a unified framework

In this section, we address RQ1. We first use the entity recognition models discussed in previous chapters to extract fine-grained knowledge entities from each paper and patent. Then, we leverage the pre-trained SciBERT model to obtain semantic vectors for the entities in both patents and papers. Next, we calculate the semantic distance between the entities to assess their novelty. The distribution of entity semantic distance-based novelty is shown in the Figure 4. The novelty score of each paper and patent is measured by the proportion of novel entities within the document.

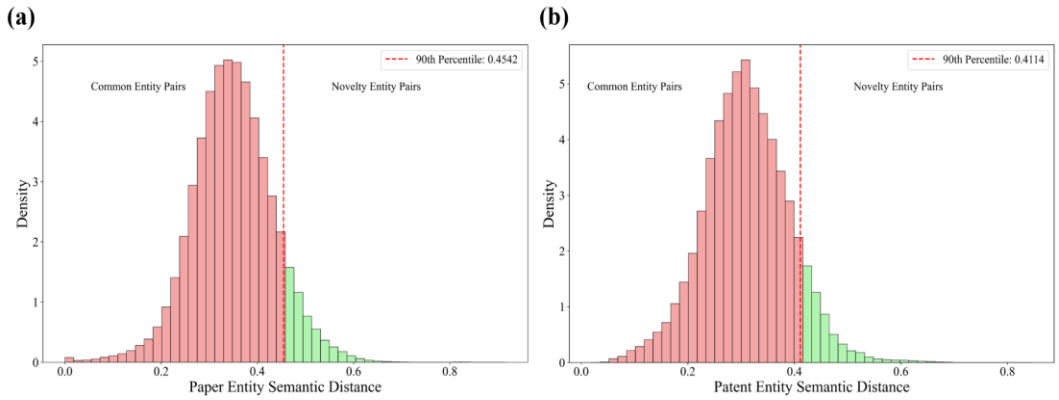


Figure 4. Semantic distance distribution of fine-grained knowledge entities (a) Semantic distance distribution of paper entities; (b) Semantic distance distribution of patent entities.

Based on the novelty of each patent and paper, we calculate the average novelty of patents and papers from each institution per year. As shown in Figure 5(a) and (b), the novelty of publications from various types of institutions in the NLP field generally exhibits an upward trend. Additionally, we observe that, the novelty trends of both papers and patents in industry are lower than those in academia and collaborations. This will be further explored in the next section.

Additionally, a six-year time window was employed, dividing the data into four intervals to assess differences over time, as shown in Figures 6. We conducted t-tests across different intervals to analyze the differences in novelty over time.

Although both paper and patent novelty trends exhibit upward growth, t the increase in novelty was more pronounced in the most recent time window (2018–2022) for patents, reflecting the rapid advancement of technological accumulation and application innovation. Although the t-tests in Figures 6(d) and 6(f) were not significant, this result is primarily due to the small number of patents related to academia and collaboration types. In contrast, the increase in novelty for NLP papers over the past six years was not significant. Several factors may contribute to this trend. First, this may be due to the gradual maturation of methodologies. Recent pre-trained models, in particular, show strong theoretical connections with earlier deep learning techniques. Second, the novelty measurement is based on the semantic distance calculated by SciBERT, whose training corpus primarily consists

of Semantic Scholar papers before 2019. Consequently, it may have limited capacity to express fine-grained knowledge entities that appear in recent papers.

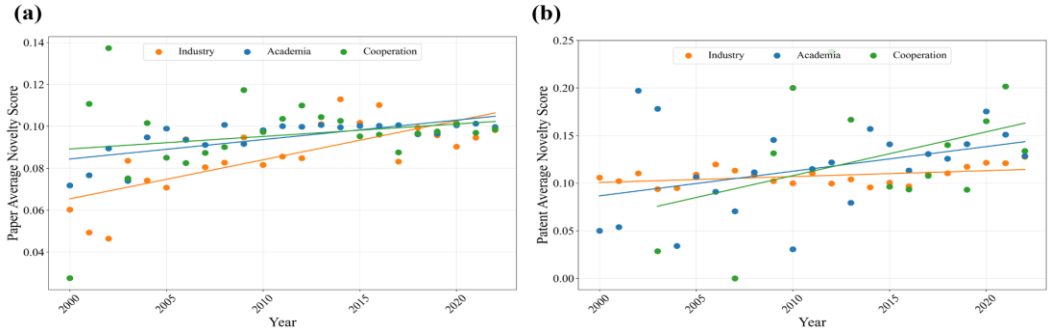


Figure 5. Trends in novelty changes of papers and patents (a) Average novelty of papers from different institutions (b) Average novelty of patents from different institutions.

Furthermore, in terms of collaboration types, Figures 6(c) and 6(f) exhibit different patterns. For papers, the novelty of collaboration types remains nearly constant across each window, while for patents, the novelty of collaboration types shows an upward trend. Although statistically insignificant (due to the small sample size). This highlights the different performances of collaboration types institutions in terms of patent and paper novelty. When it comes to industry and academia, we did not observe any significant differences in trends.

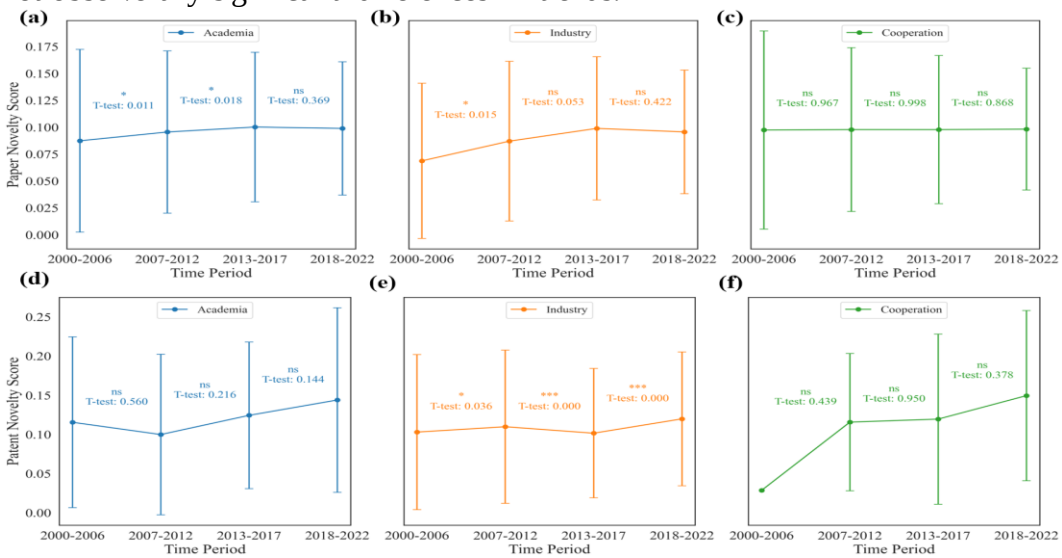


Figure 6. The differences in novelty across different time windows. (a) Novelty variation in academic papers over 6-year windows (b) Novelty variation in industry papers over 6-year windows (c) Novelty variation in cooperation papers over 6-year windows (d) Novelty variation in academic patents over 6-year windows (e) Novelty variation in industry patents over 6-year windows (f) Novelty variation in cooperation patents over 6-year windows.

Regression analysis of novelty differences across various type institutions

In this section, we focus on answering RQ2. Our preliminary analysis reveals the disparities in novelty among various types of institutions within both papers and patents, as shown in Figure 7. It is observed that academic and collaborative institutions exhibited higher novelty than industrial ones. Further, using t-tests, we found that the novelty differences between academic and collaborative institutions were not significant, with both exhibiting higher novelty than the industrial sector. To more accurately characterize the results and their reliability, we conducted regression analysis, controlling for year and institution count, to evaluate the novelty of different types of literature.

Further, we use institution type as the independent variable and introduce a series of control variables to explore the differences in novelty across different institution combinations. The regression results are shown in Tables 4 and 5.

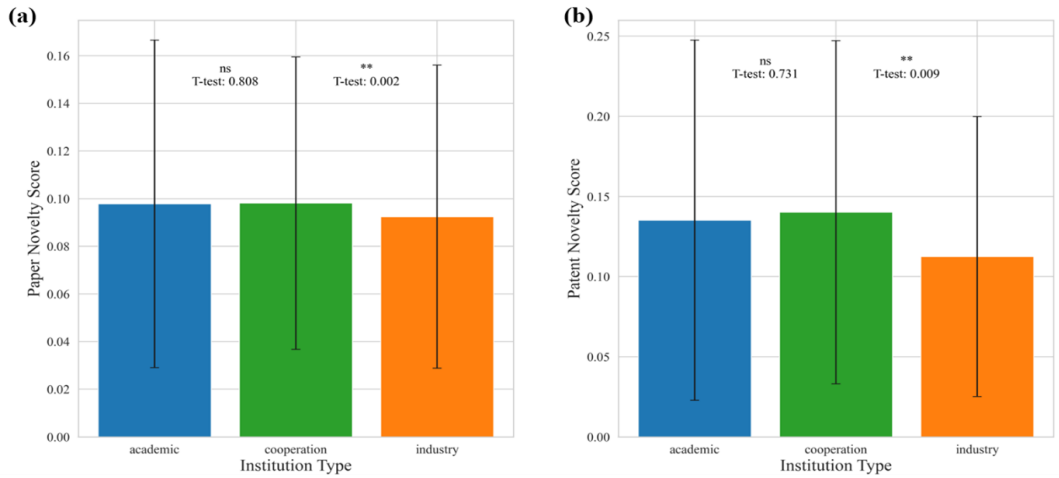


Figure 7. Box plot of novelty distribution. (a) Novelty differences across publishing institutions in the papers (b) Novelty differences across publishing institutions in the patents.

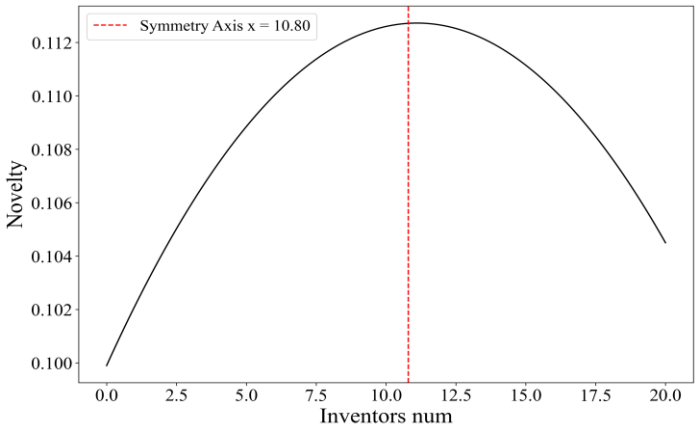


Figure 8. The relationship between the number of patent inventors and novelty. (The dashed line represents the axis of symmetry of the inverted U-shaped curve).

In the regression analysis, this study particularly focuses on the novelty performance of academia and industry in scientific papers and patents. To ensure a more focused analysis, other types of institutions were excluded. For patents, we controlled for year and institution type fixed effects, while also introducing various control variables to examine the relationship between institution type and novelty scores.

As shown in Table 5, Model (1), which includes only the independent variables, demonstrates that patents produced by academic and collaborative institutions exhibit significantly higher levels of novelty compared to those from industrial institutions. These differences are statistically significant at the 1% and 5% levels, respectively. Models (3) and (4) progressively incorporate control variables, yet the positive association between academic and collaborative institutions and patent novelty remains consistent and robust. This conclusion holds even after accounting for the number of inventors, IPC categories, and patent family size. Model (2) serves as the baseline model, exploring the relationship between team size (number of inventors) and patent novelty. The analysis reveals that the number of inventors is generally positively correlated with novelty, exhibiting a slight inverted U-shaped trend. The squared term of the number of inventors has a small but significant effect (β = 0.0001, p < 0.1). Further exploration confirms an inverted U-shaped relationship between team size and novelty. Figure 8 illustrates the trends in novelty as a function of inventor team size.

The regression analysis results at the paper level, presented in Table 6, reveal that in Model (1), which includes only the independent variables, indicates that academic papers and collaborative papers generally exhibit higher novelty. However, when the number of institutions is introduced as a control variable in Model (3), the novelty advantage of collaborative papers over industrial papers becomes statistically insignificant, suggesting that institutional factors mediate the observed effects of collaboration. Model (2) assesses the impact of the number of authors, revealing no significant correlation between the number of authors and paper novelty, in contrast to the notable role of inventor count in patents. Finally, Model (4), which includes all control variables, confirms the earlier conclusions: academic papers are still more novel compared to industrial papers, and the novelty of collaborative papers aligns more closely with that of industrial papers.

Table 5. Regression results for patent novelty.

Novelty / Variables	(1) *Model 1*	(2) *Model 2*	(3) *Model 3*	(4) *Model 4*
Academic	0.020***		0.020***	0.021***
	(0.004)		(0.004)	(0.004)
Cooperation	0.025**		0.024**	0.0245**
	(0.011)		(0.011)	(0.011)
Family size				-0.001***
				(0.000)
Inventors num		0.002***		0.002***
		(0.001)		(0.001)

Inventors num sq		-0.000*		-0.000*
		(0.000)		(0.000)
Institutions num			-0.000	-0.001
			(0.001)	(0.001)
IPC num				0.001**
				(0.001)
Constant	0.104***	0.100***	0.105***	0.102***
	(0.006)	(0.006)	(0.006)	(0.009)
Year Fixed	Yes	Yes	Yes	Yes
Observations	22269	22,269	22,269	22,269
R-squared	0.015	0.014	0.015	0.017

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 6. Regression results for paper novelty.

Novelty / Variables	(1) *Model 1*	(2) *Model 2*	(3) *Model 3*	(4) *Model 4*
Academic	0.005***		0.0054***	0.004***
	(0.002)		(0.002)	(0.002)
Cooperation	0.004**		0.0031	0.003
	(0.002)		(0.002)	(0.002)
Authors num		0.000		-0.000
		(0.000)		(0.000)
Institutions num			0.001	0.001
			(0.000)	(0.000)
Constant	0.063***	0.067***	0.063***	0.063***
	(0.007)	(0.006)	(0.007)	(0.007)
Year Fixed	Yes	Yes	Yes	Yes
Observations	17,644	17,644	17644	17,644
R-squared	0.005	0.004	0.0056	0.005

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Robustness checks

We conducted a robustness check on the previous results to verify the reliability of the findings. Furthermore, we binarized the novelty scores by labeling the top 10% of papers with the highest novelty scores as "novel," while the remaining papers were labeled as "non-novel" (Jeon et al., 2022). Subsequently, we reanalyzed the data using logistic regression, and the results, as shown in Tables 7 and 8, were consistent with the previous findings.

Table 7. Regression results of patent novelty with novelty as a binary variable.

Novelty / Variables	(1) *Model 1*	(2) *Model 2*	(3) *Model 3*	(4) *Model 4*
Academic	0.587***		0.587***	0.612***
	(0.127)		(0.127)	(0.127)
Cooperation	1.015***		1.02***	1.023***
	(0.287)		(0.295)	(0.297)

Family size				-0.022
				(0.013)
Inventors num		0.073***		0.070***
		(0.026)		(0.026)
Inventors num sq		-0.004		-0.003
		(0.002)		(0.002)
Institutions num			-0.006	-0.047
			(0.055)	(0.05)
IPC num				-0.003
				(0.024)
Constant	-2.171***	-2.313***	-2.164***	-2.101***
	(0.220)	(0.225)	(0.228)	(0.332)
Year Fixed	Yes	Yes	Yes	Yes
Pseudo R-squared	0.002	0.001	0.002	0.003

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 8. Regression results of paper novelty with novelty as a binary variable.

Novelty / **Variables**	(1) *Model 1*	(2) *Model 2*	(3) *Model 3*	(4) *Model 4*
Academic	0.256***		0.236**	0.221**
	(0.094)		(0.094)	(0.095)
Cooperation	0.120		0.04	0.027
	(0.103)		(0.019)	(0.110)
Authors num		0.006		-0.017
		(0.012)		(0.014)
Institutions num			0.046**	0.063**
			(0.020)	(0.026)
Constant	-2.469***	-2.259***	-2.510***	-2.483***
	(0.341)	(0.333)	(0.341)	(0.342)
Year Fixed	Yes	Yes	Yes	Yes
Pseudo R-squared	0.001	0.000	0.001	0.002

Note: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Discussion

This study adopted fine-grained knowledge entity analysis to evaluate the novelty of patents and papers within the NLP field. Based on previous entity-based novelty metrics, we further optimized the novelty measurement method. Through regression analysis, it was revealed that new ideas in the NLP field are continuously emerging (Zhang et al., 2024). Moreover, the level of novelty in academia surpasses that in industry when considering both papers and patents. This finding is consistent with the results of (Chen et al., 2024; Liang et al., 2024). Further research has found that, at the paper level, academic–industry collaborations struggle to replicate the novelty of academic teams and tend to resemble the work of industry teams (Liang et al., 2024).

As a catalyst, academia significantly promotes the enhancement of novelty, both in terms of filing patents individually and participating in patent research composition, thereby enabling the industry to disclose more innovative findings. This trend remains significant after controlling for the number of institutions and other relevant variables. This not only helps advance patent technologies to higher levels but also provides more competitive technological solutions for the industry. As Krieger et al. (2024) point out, scientific research enables companies to derive significantly more value from their inventions, and patents closer to science tend to exhibit higher novelty. In contrast, at the paper level, although academia overall performs with greater novelty, after controlling for the number of institutions, the impact of collaboration type and team size on the novelty of scientific papers is relatively small. This study only found an inverted U-shaped relationship between the size of collaborative teams and novelty in patents.

Implications

Theoretical implications: The theoretical significance of this study is reflected in the following three aspects: First, by transferring the paper entity recognition model knowledge to the patent entity recognition model and combining it with an entity-based novelty measurement method, this study achieves a unified measurement of novelty in both patents and papers. This provides a feasible framework for evaluating the novelty of paper and patent levels across a broader dataset. Second, this study provides new empirical evidence, revealing the novelty differences between academia and industry in the NLP field, both in patents and scientific papers, and highlights how novelty varies across different types of institutions. Finally, this study examines the relationship between team characteristics and novelty in the NLP field, particularly how team size impacts the novelty of research outcomes. It confirms that larger inventor teams, by combining diverse expertise, tend to innovate within familiar knowledge domains (Uzzi et al., 2013). However, when team size exceeds a certain threshold, increased coordination costs and communication challenges lead to incremental improvements rather than novel breakthroughs. This suggests that larger teams in the patent field may experience reduced innovation novelty, relying more on established solutions (Wu et al., 2019; Shin et al., 2022). This finding contributes to the understanding of research team formation and collaboration models in the NLP field.

Practical implications: The results of this study offer theoretical support for the distinct roles of academia and industry in technological innovation, while providing practical recommendations for optimizing research team composition and size. The findings show that academia generally exhibits higher novelty in both patents and papers, highlighting the importance of academic institutions' role in advancing fundamental research and innovation. Academia's openness and collaboration foster new ideas and support interdisciplinary efforts (Brescia et al., 2014). This study also reveals the impact of team size on novelty. In technology-intensive fields like NLP, larger inventor teams can drive innovation by integrating diverse expertise. While reasonable team size and interdisciplinary collaboration foster breakthroughs, overly large teams may increase coordination costs and dilute focus,

reducing innovation efficiency. According to the regression analysis, the "threshold" for inventor teams appears to be around 10 members. For typical inventor teams, increasing team size helps improve patent novelty. However, for scientific papers, the number of authors does not directly affect innovation, indicating that novelty depends more on research depth and collaboration model than on team size or cross-sector collaboration work.

Limitations

Despite adjustments to the entity-based novelty measurement method and empirical analysis revealing novelty differences between academia and industry, this study has some limitations. First, while we classified entity relationships and quantified semantic distances, the removal of specific entity types remains coarse. Future research should refine entity distance measurements, especially for same-type and different-type entities, or incorporate discourse structure information. Additionally, there is some discrepancy between semantic distance and the difficulty of combining fine-grained knowledge entities. Future studies could explore combining graph representation learning with co-occurrence network topology to improve novelty assessment. Finally, although this study's dataset covers a wide range of patents and papers, the sample size in the NLP field is relatively limited. Additionally, the imbalanced distribution of institutions in the paper and patent data, especially in the patent data, may potentially affect the accuracy of the analysis results. In addition, although we found that the novelty of industry outputs is lower than that of academia, we did not further explore the reasons behind this. The study did not address whether the disclosure strategy of industry is more conservative, or if the research content itself lacks sufficient novelty. Finally, while we included several key factors that are easy to capture and control in the regression, other variables may have been overlooked, potentially influencing the study's outcomes.

Conclusion and future works

This study explores novelty differences between academia and industry. By extracting fine - grained knowledge entities and measuring paper novelty based on novel entity proportions, regression models analyse novelty differences in patents and papers from academia and industry.

Results show academia has a novelty advantage in both patents and papers, especially in patents. In scientific papers, the impact of collaboration type on novelty is insignificant when controlling for team size. There's an inverted U - shaped relationship between patent team size and novelty in the NLP field. For scientific papers with small inventor teams, increasing team size and cross - disciplinary collaboration can boost patent novelty.

Future research directions include: expanding the sample to the AI field to validate findings; using graph representation learning and entity connection frequency, instead of just semantic distance, to measure novelty; and exploring the mechanisms behind the greater patent novelty in academia - industry collaboration by examining factors like scientific - technical distance, institutional research backgrounds, and disclosure strategies.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Grant No.72074113).

References

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, 379(6635), 884–886.

Ahuja, G., & Lampert, C. M. (2001). Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, *22*(6–7), 521–543.

Arinas, I. (2012). How vague can your patent be? Vagueness strategies in US patents. Vagueness Strategies in US Patents. *HERMES-Journal of Language and Communication in Business*, *48*, 55-74.

Arts, S., Hou, J., & Gomez, J. C. (2019). Text Mining to Measure Novelty and Diffusion of Technological Inventions. In *Proceedings of the 1st Workshop on Patent Text Mining and Semantic Technologies*. Karlsruhe, Germany. https://doi.org/10.34726/pst2019.2

Arts, S., Hou, J., & Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, *50*(2), 104144.

Arundel, A. (2001). The relative effectiveness of patents and secrecy for appropriation. *Research Policy*, *30*(4), 611–624.

Ashtor, J. H. (2021). Modeling patent clarity. *Research Policy*, *51*(2), 104415.

Anne, K. (2023). Data-drivenness, novelty, and interdisciplinarity in the study of criminology. In *Proceedings of the International Society for Scientometrics and Informetrics*, 2, 225–231. https://doi.org/10.5281/zenodo.8370929

Ba, Z., Meng, K., Ma, Y., & Xia, Y. (2024). Discovering technological opportunities by identifying dynamic structure-coupling patterns and lead-lag distance between science and technology. *Technological Forecasting & Social Change,* 200(6351), Article 123147.

Bikard, M., & Marx, M. (2019). Bridging Academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, *66*(8), 3425–3443.

Brescia, F., Colombo, G., & Landoni, P. (2016). Organizational structures of Knowledge Transfer Offices: an analysis of the world's top-ranked universities. *The Journal of Technology Transfer*, 41(1), 132–151.

Chan, H. F., & Torgler, B. (2020). Gender differences in performance of top cited scientists by field and country. *Scientometrics*, *125*(3), 2421–2447.

Chen, C., Hsiao, Y., Chu, M., & Hu, K. (2015). The relationship between team diversity and new product performance: the moderating role of organizational slack. *IEEE Transactions on Engineering Management*, *62*(4), 568–577.

Chen, Z., Zhang, C., Zhang, H., Zhao, Y., Yang, C., & Yang, Y. (2024). Exploring the relationship between team institutional composition and novelty in academic papers based on fine-grained knowledge entities. *The Electronic Library*, *42*(6): 905-930..

Chirico, F., Criaco, G., Baù, M., Naldi, L., Gomez-Mejia, L. R., & Kotlar, J. (2018). To patent or not to patent: That is the question. Intellectual property protection in family firms. *Entrepreneurship Theory and Practice*, *44*(2), 339–367.

Clarysse, B., Andries, P., Boone, S., & Roelandt, J. (2023). Institutional logics and founders' identity orientation: Why academic entrepreneurs aspire lower venture growth. Research Policy, 52(3), Article 104713.

Codina-Filbà, J., Bouayad-Agha, N., Burga, A., Casamayor, G., Mille, S., Müller, A., Saggion, H., & Wanner, L. (2016). Using genre-specific features for patent summaries. *Information Processing & Management*, *53*(1), 151–174.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. (pp. 4171–4186). Minneapolis, Minnesota.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., . . . Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 57(2021), Article 101994.

Färber, M., & Tampakis, L. (2023). Analyzing the impact of companies on AI research based on publications. Scientometrics, 129(1), 31-63.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, *47*(1), 117–132.

Geisler, E. (1995). When whales are cast ashore: the conversion to relevancy of American universities and basic science. *IEEE Transactions on Engineering Management*, *42*(1), 3–8.

Homscheid, D., Kunegis, J., & Schaarschmidt, M. (2015). Private-Collective Innovation and Open Source Software: Longitudinal Insights from Linux Kernel Development. In *Conference on e-Business, e-Services and e-Society (pp. 299-313)*. (pp. 299–313).

Hottenrott, H., Rose, M. E., & Lawson, C. (2021). The rise of multiple institutional affiliations in academia. Journal of the Association for Information Science and Technology, 72(8), 1039–1058.

Jang, H., Kim, S., & Yoon, B. (2023). An eXplainable AI (XAI) model for text-based patent novelty analysis. *Expert Systems Withwith Applications*, *231*, 120839.

Jee, S. J., & Sohn, S. Y. (2023). Firms' influence on the evolution of published knowledge when a science-related technology emerges: the case of artificial intelligence. *Journal of Evolutionary Economics*, 33(1), 209–247.

Jeon, D., Ahn, J. M., Kim, J., & Lee, C. (2022). A doc2vec and local outlier factor approach to measuring the novelty of patents. *Technological Forecasting and Social Change*, *174*, 121294.

Jeon, D., Lee, J., Ahn, J. M., & Lee, C. (2023). Measuring the novelty of scientific publications: A fastText and local outlier factor approach. *Journal of Informetrics*, *17*(4), 101450.

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021). Natural Language Processing: history, evolution, application, and future work. In *Proceedings of 3rd International Conference on Computing Informatics and Networks*. (pp. 365–375). Delhi, India.

Kabore, F. P., & Park, W. G. (2019). Can patent family size and composition signal patent value? *Applied Economics*, *51*(60), 6476–6496.

Koppman, S., & Leahey, E. (2019). Who moves to the methodological edge? Factors that encourage scientists to use unconventional methods. *Research Policy*, *48*(9), 103807.

Krieger, J. L., Schnitzer, M., & Watzinger, M. (2024). Standing on the shoulders of science. *Strategic Management Journal*, *45*(9), 1670–1695.

Kuhn, J., Younge, K., & Marco, A. (2020). Patent citations reexamined. *The RAND Journal of Economics*, *51*(1), 109–132.

Larivière, V., Macaluso, B., Mongeon, P., Siler, K., & Sugimoto, C. R. (2018). Vanishing industries and the rising monopoly of universities in published research. *PLoS ONE*, *13*(8), e0202120.

Lee, C., & Lee, G. (2019). Technology opportunity analysis based on recombinant search: patent landscape analysis for idea generation. *Scientometrics*, *121*(2), 603–632.

Lee, Y., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*(3), 684–697.

Liang, L., Zhuang, H., Zou, J., & Acuna, D. E. (2024). The complementary contributions of academia and industry to AI research. (arVix:2401.10268). *arXiv*.

Liang, Z., Mao, J., & Li, G. (2022). Bias against scientific novelty: A prepublication perspective. *Journal of the Association for Information Science and Technology*, *74*(1), 99–114.

Liu, M., Bu, Y., Chen, C., Xu, J., Li, D., Leng, Y., . . . Ding, Y. (2022). Pandemics are catalysts of scientific novelty: Evidence from COVID-19. *Journal of the Association for Information Science and Technology*, *73*(8), 1065–1078.

Liu, M., Xie, Z., Yang, A. J., Yu, C., Xu, J., Ding, Y., & Bu, Y. (2024). The prominent and heterogeneous gender disparities in scientific novelty: Evidence from biomedical doctoral theses. *Information Processing & Management*, *61*(4), 103743.

Luo, Z., Lu, W., He, J., & Wang, Y. (2022). Combination of research questions and methods: A new measurement of scientific novelty. *Journal of Informetrics*, *16*(2), 101282.

MacRoberts, M. H., & MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, *36*(3), 435–444.

Marcoulides, K. M., & Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable modeling methods. *Educational and Psychological Measurement*, *79*(5), 874–882.

Martinez-Senra, A. I., Quintas, M. A., Sartal, A., & Vazquez, X. H. (2015). How Can Firms' Basic Research Turn Into Product Innovation? The Role of Absorptive Capacity and Industry Appropriability. *IEEE Transactions on Engineering Management*, *62*(2), 205–216.

Mori, T., & Sakaguchi, S. (2018). Collaborative knowledge creation: Evidence from Japanese patent data. (arXiv: 1908.01256). *arXiv*.

Nelson, R., & Winter, S. (1982). *An evolutionary theory of economic change*. Harvard University Press.

Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*(7942), 138–144.

Perkmann, M., & Walsh, K. (2009). The two faces of collaboration: impacts of university-industry relations on public research. *Industrial and Corporate Change*, *18*(6), 1033–1065.

Pramanick, A., Hou, Y., Mohammad, S. M., & Gurevych, I. (2024). The Nature of NLP: Analyzing contributions in NLP papers. (arXiv: 2409.19505). *arXiv*.

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. *Preprint*. 1–12.

Riera, R., & Rodríguez, R. (2022). What if Peer-Review process is killing Thinking-Out-of-the-Box science? *Frontiers in Marine Science*, 9, Article 924469.

Sauermann, H., & Stephan, P. E. (2010). Twins or strangers? Differences and similarities between industrial and academic science. *National Bureau of Economic Research*, (No. w16113).

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. *PLoS ONE*, *16*(7), e0254034.

Shin, H., Kim, K., & Kogler, D. F. (2022). Scientific collaboration, research funding, and novelty in scientific knowledge. *PLoS ONE*, *17*(7), e0271678.

Sun, X., Chen, N., & Ding, K. (2022). Measuring latent combinational novelty of technology. *Expert Systems With Applications*, *210*, 118564.

Tao, A., Qi, Q., Li, Y., Da, D., Boamah, V., & Tang, D. (2022). Game Analysis of the Open-Source Innovation Benefits of Two Enterprises from the Perspective of Product Homogenization and the Enterprise Strength Gap. *Sustainability*, *14*(9), 5572.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157), 468–472.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need (arXiv:1706.03762). *arXiv*.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436.

Wei, T., Feng, D., Song, S., & Zhang, C. (2024). An extraction and novelty evaluation framework for technology knowledge elements of patents. *Scientometrics*.

Wu, K., Xie, Z., & Du, J. T. (2024). Does science disrupt technology? Examining science intensity, novelty, and recency through patent-paper citations in the pharmaceutical field. *Scientometrics*.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382.

Wu, M., Chang, K., Zhou, W., Hao, J., Yuan, C., & Chang, K. (2015d). Patent deployment Strategies and patent Value in LED industry. *PLoS ONE*, *10*(6), e0129911.

Xu, H., Bu, Y., Liu, M., Zhang, C., Sun, M., Zhang, Y., Meyer, E., Salas, E., & Ding, Y. (2022). Team power dynamics and team impact: New perspectives on scientific collaboration using career age as a proxy for team power. *Journal of the Association for Information Science and Technology*, *73*(10), 1489–1505.

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences*, *119*(36), e2200841119.

Zanella, G., Liu, C. Z., & Choo, K. R. (2021). Understanding the trends in blockchain domain through an unsupervised systematic patent analysis. *IEEE Transactions on Engineering Management*, *70*(6), 1991–2005.

Zhang, H., Zhang, C., & Wang, Y. (2024). Revealing the technology development of natural language processing: A Scientific entity-centric perspective. *Information Processing & Management*, *61*(1), 103574.

Zheng, H., Li, W., & Wang, D. (2022). Expertise diversity of teams predicts originality and Long-Term impact in science and technology. (arXiv: 2210.04422). *arXiv*.

https://doi.org/10.51408/issi2025_036

Exploring Scientist's Research Trajectories within a Field with Main Path Analysis

Chung-Huei Kuan[1], Szu-Chia Lo[2]

[1]*maxkuan@mail.ntust.edu.tw*
National Taiwan University of Science and Technology, Graduate Institute of Patent, Taipei (Taiwan, ROC)

[2]*szuchialo@ntu.edu.tw*
National Taiwan University, Department of Library and Information Science, Taipei (Taiwan, ROC)

Abstract

Main Path Analysis (MPA) is commonly applied to citation networks constructed from papers or patents within a research or technological field to reveal representative knowledge-diffusion trajectories for that field. These trajectories, known as Main Paths (MPs), reflect the overall knowledge development and evolution within the field.

Rather than examining field-level trajectories, this study introduces a novel approach to explore individual scientists' research trajectories within the field. These individual-level trajectories enable analysts to trace the lineage of a scientist's work, understand its origins, and uncover its influence on subsequent research. Additionally, these individual-level trajectories can be contrasted with field-level trajectories to examine their interactions, providing further insights into a scientist's contributions relative to the field's mainstream development.

This approach relies on a previously overlooked path search algorithm in MPA, referred to as key-node search, to generate MPs that capture distinct knowledge flows centered around a scientist's works. A case study based on patents in the field of Evolutionary Computation, using an official artificial intelligence patent dataset, demonstrates both a macro-view and a micro-view of the proposed individual-level MPs.

Introduction

Hummon and Doreian (1989) developed the so-called Main Path Analysis (MPA), which aims to uncover "the mainstream of literature of a clearly delineated area of scientific research" from the citation network of a specific research area. Since its inception, MPA has become a widely recognized method, leading to a proliferation of studies employing it. Its popularity can be attributed to its conceptual simplicity, further bolstered by its availability in the popular network analysis tool Pajek (Batagelj & Mrvar, 1998; De Nooy, Mrvar, & Batagelj, 2018).

MPA is typically employed to analyze networks of mutually citing documents, such as scientific papers or patent publications associated with a specific field of study. In such networks, documents are represented as nodes, while their citations are represented as arcs, denoting pathways for the flow of knowledge from cited documents to citing ones. By applying MPA, one or more series of connected arcs—referred to as main paths (MPs)—are derived from the network and identified as representative trajectories of knowledge development within the field.

Rather than focusing solely on the MPs for a field, or field-level MPs, this study explores whether the same approach can be applied to individual researchers or

scientists within the field. Specifically, it investigates whether the most representative trajectories passing through the papers or patents associated with a researcher or scientist can be identified as individual-level MPs. To the best of the authors' knowledge, no prior study has explored this endeavor of uncovering individual-level trajectories. Therefore, this study aims to fill that gap.

Uncovering individual-level MPs offers several benefits. First, these MPs can illuminate how a researcher's or scientist's works evolve, particularly how they are influenced by or contribute to other works in the field. Second, these MPs can be compared against those of other researchers or scientists to explore how their trajectories interrelate. Their respective MPs may run parallel, diverge, or converge at certain points, revealing overlaps or intersections in their research efforts. Third, these MPs can also be contrasted with field-level MPs to examine their interactions, offering deeper insights into a researcher's or scientist's contributions relative to the field's mainstream development.

Literature Review

Overview of MPA

To derive MPs from a citation network, MPA primarily involves two key components. First, weights are assigned to the network's arcs to reflect their significance in knowledge diffusion (Hummon & Doreian, 1989). Once the arc weights are assigned, MPA performs a path search on the weighted network to identify chains of connected arcs that extend from the network's sources to its sinks, which are then identified as MPs.

There are various weight assignment and path search algorithms in MPA. The most widely used weight assignment algorithms—namely SPC, SPLC, and SPNP—are collectively known as SPX algorithms. For an in-depth description of these algorithms, refer to Kuan (2020).

The most popular path search algorithms, such as those available in Pajek, can be broadly categorized into global and local searches, each with a number of similar variants listed in Table 1. The approach introduced in this study is based on a path search algorithm called the key-node search (Kuan, 2024; Kuan & Liao, 2024), which also has global and local variants (more details are provided later).

Table 1. Categorization of common path search algorithms.

Category	***Variants***	***Related parameters***
Global searches	Global standard	
	Global key-route	arcs having the topmost *N* weights as key routes
	Global key-node	a designated set of key nodes
Local searches	Local forward	a tolerance value between 0 and 1
	Local backward	a tolerance value between 0 and 1
	Local key-route	1) a tolerance value between 0 and 1 2) arcs having the topmost *N* weights as key routes
	Local key-node	1) a tolerance value between 0 and 1 2) a designated set of key nodes

Global searches identify MPs by selecting paths having the highest path weights (i.e., the sum of all arc weights along a path) between two sets of nodes. In contrast, local searches construct the MPs incrementally, progressing step by step from one set of nodes to another.

More specifically, global standard (GS) search (Liu & Lu, 2012) selects MPs from the paths between the network's sources and sinks. Local forward (LF) search (Hummon & Doreian, 1989) begins at the sources and progressively selects the highest weighted outgoing arcs, moving to subsequent nodes until a sink is reached. Conversely, local backward (LB) search (Hummon & Doreian, 1989) starts at the sinks and traces backward through the highest weighted incident arcs until a source is reached. When conducting local search, a tolerance value can be set to include arcs within a specified range of the highest weight for tracing (De Nooy, Mrvar, & Batagelj, 2018). For instance, a local search with the tolerance value of 0.10 would trace incident or outgoing arcs with weights that are at least 90% of the highest weight among them.

Key-route searches (Liu & Lu, 2012) develop MPs by starting with a set of highest-weighted arcs, referred to as key routes. For a given key route (*i*, *j*), the global key-route (GKR) search performs the GS searches between the sources and the arc's start node *i*, and between the arc's end nodes *j* and the sinks, to identify one or more global paths preceding and succeeding the key route (*i*, *j*), respectively. These global paths are then concatenated with the key route (*i*, *j*) to form its GKR MPs. Similarly, the local key-route (LKR) search employs LB and LF searches, instead of GS searches, to derive the preceding and succeeding paths for the key route (*i*, *j*).

Key-route searches involve a parameter *N*, which specifies arcs with the topmost *N* weights to be used as key routes. For instance, a key-route 1 search initiates MP development from the arc with the highest weight, while a key-route 10 search includes arcs with weights up to the 10th highest. In key-route *N* searches, the number of key routes may exceed *N* if weights are tied.

Key individuals along the MPs

There is a wealth of research involving the application of MPA to uncover a field's field-level MPs. Among these studies, some have also focused on identifying significant individuals, especially firms, within the field. These studies generally follow a common approach: they first derive the field-level MPs and then identify individuals whose works appear on these MPs. Such individuals are considered key contributors, as their works are integral to the most representative trajectories of the field's evolution.

Recent studies provide several examples. For instance, Su, Chen, Chang, and Lai (2019) employed MPA to trace the dominant knowledge flow in the field of blockchain technology and identified owners of the patents on the MPs as key players for the field. The study then analyzed the patent families of these key players to investigate their strategic intent in managing their patent portfolios.

Cho, Liu, and Ho (2021) applied MPA to patents related to autonomous driving to uncover the technology development trajectory for the field. The study identified assignees whose patents appeared on the trajectory as key players. Additionally,

based on the different phases along the development trajectory and the associated key assignees within each phase, the study categorized these key players into groups such as "technology developers," "technology integrators," and "technology implementers." A similar methodology was adopted by Chen and Cho (2023) to analyze trends and identify key players in the field of Low Earth Orbit (LEO) satellite technology using patents.

Watanabe and Takagi (2021) used MPA to examine how technology has evolved within the field of computer graphic processing systems. The study developed MPs for the field at 5-year intervals and observed the appearance and disappearance of firms owning patents on the MPs over time. The authors noted that these patterns of firm appearances and disappearances align with the historical evolution of the field.

The above studies have several limitations. Firstly, as only individuals with works along the MPs are considered, those without any works on the MPs are overlooked. Additionally, field-level MPs may fail to capture other relevant or even key individuals, as Verspagen (2007) empirically demonstrated that MPA is highly selective at the firm level, with many active individuals in the field not present on the MPs. Furthermore, the identified individuals may have additional works beyond those located on the field-level MPs, which may be overloooked under this approach.

MPs from specific nodes

The key-node search algorithms employed in this study is similar to the key-route algorithms, with the key distinction being that they begin MP development from a set of analyst-designated key nodes, rather than a number of top-weighted key routes determined for the analyst. More details on this approach will be provided in the Methodology section.

This study has identified several prior works with methodologies akin to the approach adopted here. Unlike traditional MPA, which typically develops field-level MPs by searching the citation network from sources to sinks, or vice versa (except for the key-route searches described earlier), these studies first analytically identified a number of key documents. They then developed field-level MPs starting from the nodes of these key documents.

Park and Magee (2017, 2019) introduced a modified MPA that develops field-level MPs from designated nodes. The authors first identified patents with high knowledge persistence—a measure of the extent to which knowledge remains in the patents or contributes to later patents based on their structural positions in the patent citation network. They then developed field-level MPs exclusively from the nodes of these so-called high-persistence patents (HPPs), tracing forward to the sinks and backward to the sources. Feng and Magee (2020) followed a similar approach in analyzing patents from four domains of electric vehicles. They derived MPs for each domain from a number of HPPs and identified the assignees of these HPPs as key players for each domain.

Unlike the above studies, which analytically selected key nodes from the citation network, this study manually designates nodes representing the works of a specific researcher or scientist as key nodes. The resulting key-node MPs are therefore referred to as the researcher's or scientist's individual-level MPs. As these MPs are

constructed from the field's citation network, rather than using the researcher's or scientist's works in isolation, they reflect how their works evolve within the broader context of the field to which they belong.

In addition to the above-mentioned studies, several works have also explored path development from designated nodes. Ho, Saw, Lu, and Liu (2014) developed a method called "branch paths" to address the risk that minor technologies may be overshadowed by more prominent technologies and thus omitted from the field-level MPs. This method identifies a set of documents related to these minor technologies and traces paths from these designated documents both forward and backward until they encounter a node on the field-level MPs. Liu, Lu, and Ho (2019) referred to this method as the "designated-document approach" and suggested that it could reveal the relationship between these designated documents and the field-level MPs.

While these works also develop paths from specific nodes, their aim is to supplement field-level MPs rather than derive MPs from the perspective of individual researchers or scientists.

Methodology

Key-node search

As mentioned earlier, the key-node search includes global and local variants, similar to the global key-route (GKR) and local key-route (LKR) searches, as summarized in Table 1. The primary distinction is that key-node MPs are derived from specific nodes that are manually designated as key nodes by the analyst. In contrast, in the key-route search, the analyst cannot specify individual arcs as key routes but can only control the parameter N.

As illustrated in Figure 1, for a designated key node k (the white node), the key-node search identifies the representative preceding and succeeding paths (depicted in solid lines) between the sources (dark nodes to the left) and the key node k, and between k and the sinks (dark nodes to the right). These paths may pass through intermediate nodes (gray nodes). In global key-node (GKN) search, the representative preceding and succeeding paths are derived using global standard (GS) search, whereas in the local key-node (LKN) search, they are determined using local backward (LB) and local forward (LF) searches, respectively. These representative preceding and succeeding paths are then cascaded to form the MPs for the key node k. Finally, the MPs for all key nodes are aggregated to form the overall key-node MPs.

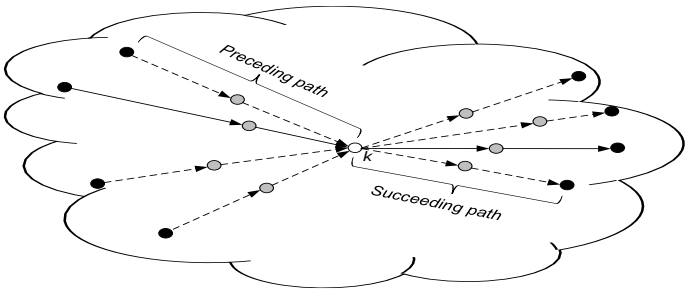


Figure 1. MPs by Key-node search.

In other words, the key-node search constructs MPs by initiating the development of significant paths both preceding and succeeding the designated key nodes. By assigning nodes that represent a researcher's or scientist's works as key nodes, the resulting key-node—or researcher's or scientist's individual-level—MPs reveal, on one hand, the works within the broader field that have most influenced the researcher's or scientist's works and, on the other hand, the representative subsequent developments stemming from the researcher's or scientist's works, also within the broader context of the field.

The individual-level MPs, therefore, offer deeper insights into the research evolution of researchers or scientists than simply aligning their works chronologically. Furthermore, individual-level MPs facilitate a more nuanced understanding of the interrelationships among the researcher's or scientist's works. For instance, some works may appear on separate paths within the individual-level MPs, suggesting that they stem from distinct developmental trajectories in the researcher's or scientist's intellectual endeavors. Conversely, instances where multiple works appear on the same path indicate a continuation of research efforts, signifying progressive knowledge expansion within a single thematic or methodological direction.

Applications of key-node search

Like the common path searches mentioned earlier, the key-node search described above is also available in Pajek, making it readily accessible to analysts. However, perhaps due to its introduction only after 2018—where it is obscurely labeled as "through vertices in cluster"—this path search has seen little application in the literature. To promote awareness of this method and to better reflect its characteristics and similarity to the key-route search, the term "key-node search" has been coined.

Despite its simplicity, the key-node search has the potential to enhance MPA in ways that other common path searches do not. Based on the few related studies available, the following are two examples of its potential applications.

One challenge in MPA is the lack of a quantitative measure to evaluate how well MPs accurately capture and reflect overall knowledge development within a field. To address this, Kuan and Liao (2024) proposed that the representativeness of MPs is limited to the portions of the network that are reachable from or to the MPs, referring to these portions as the MPs' coverage. The study further suggested that the proportion of documents falling within the MPs' coverage can serve as a quantitative measure of their representativeness.

To uncover MPs' coverage, the study applied the LKN search with a tolerance value of 1, using all nodes on the MPs as key nodes. This approach allowed the LKN search to trace all incident and outgoing arcs for each MP node, thereby encompassing the portions of the network that were reachable from or to the MP nodes. When a significant portion of the network fell outside the MPs' coverage, reflecting a low representativeness for the MPs, the study proposed a method to identify auxiliary MPs from this out-of-coverage portion. This portion was also determined using the LKN search with a tolerance value of 1, where the key nodes included those lying outside the MPs' coverage.

Kuan (2024) observed that MPA analysts often possess domain knowledge about the field under analysis, including seminal works crucial to its development. Rather than leaving this knowledge unused in the MPA process or restricting its use solely to document collection or validation of obtained MPs, the study proposed manually incorporating these seminal works into MPA using the key-node search to generate MPs that capture a distinct knowledge flow centered around these key documents.
The study further suggested observing key-document MPs alongside field-level MPs to simultaneously examine the focused knowledge flow through key documents and the overall knowledge flow of the field. This concurrent observation allows for an analysis of their interactions, providing additional insights into the field's development. To facilitate this process, the study proposed generating key-document and field-level MPs automatically and simultaneously, both using key-node searches.
While the key-node search may seem like just one of many path search options in MPA, Kuan (2024) formally verified that the field-level MPs generated by the popular path search algorithms listed in Table 1 can all be reproduced using the key-node search with appropriately selected key nodes—except for key-route MPs, which are subject to certain preconditions. This finding establishes the key-node search as a uniquely versatile method among the algorithms listed in Table 1.

Case study

Data set

To demonstrate the real-world application of the proposed approach, this study conducts a case study using the publicly available Artificial Intelligence (AI) Patent Dataset provided by the United States Patent and Trademark Office (USPTO). This dataset comprises 13,244,037 U.S. patent documents, including utility patents and pre-grant publications (PGPubs), spanning the years 1976 to 2020.
Each patent document is classified by the USPTO using a machine learning approach to predict its relevance to one of eight AI technology fields: machine learning (ML), natural language processing (NLP), computer vision (CV), speech (S), knowledge processing (KP), AI hardware (AIH), evolutionary computation (EC), and planning and control (P&C) (Giczy, Pairolero, & Toole, 2022).
This study selects patent documents predicted to belong to the field of Evolutionary Computation (EC), resulting in 48,999 patent documents covering 36,560 inventions. EC is chosen for its versatility, which makes it a foundational approach in modern AI research and applications, offering potentially diverse and complex knowledge flows for analysis.
EC draws inspiration from biological evolution to solve optimization and search problems. It encompasses a family of techniques, including genetic algorithms, genetic programming (applying a genetic algorithm to a population of computer programs), and differential evolution (generating new candidate solutions by combining the differences between randomly selected individuals in a population of candidate solutions), which simulate natural selection, mutation, crossover, and survival of the fittest to iteratively refine solutions (Bäck, Hammel, & Schwefel,

1997). EC is widely applied in machine learning, robotics, optimization, and complex system design due to its ability to efficiently explore large search spaces and adapt to dynamic environments.

As for the researcher or scientist whose research trajectory is to be observed, this study selects John R. Koza, a pioneer in the EC field (Mitchell & Taylor, 1999). He is known for his work in genetic programming, particularly in automated program generation, where computer programs are evolved to solve complex tasks. Mr. Koza is listed as the inventor on 14 U.S. patents, 12 of which are predicted to be EC-related in the AI Patent Dataset. This study considers these 12 patents to constitute Mr. Koza's body of research for analysis. These patents are listed in Table 2, arranged in ascending order of their application dates.

Table 2. Patents with John R. Koza as the sole inventor or one of the inventors.

#	*App. no.*	*App. date*	*Pub. no.*	*Pub. date*	*Title*
1	7196973	19880520	4935877	19900619	Non-linear genetic algorithms for solving problems
2	7584259	19900918	5148513	19920915	Non-linear genetic process for use with plural co-evolving populations
3	7787748	19911105	5136686	19920804	Non-linear genetic algorithms for solving problems by finding a fit composition of functions
4	7881507	19920511	5343554	19940830	Non-linear genetic process for data encoding and for solving problems using automatically defined functions
5	7899627	19920616	5390282	19950214	Process for problem solving using spontaneously emergent self-replicating and self-improving entities
6	8286134	19940804	5742738	19980421	Simultaneous evolution of the architecture of a multi-part program to solve a problem using architecture altering operations
7	8603648	19960220	5867397	19990202	Method and apparatus for automated design of complex structures using genetic programming
8	8813894	19970307	6058385	20000502	Simultaneous evolution of the architecture of a multi-part program while solving a problem using architecture altering operations
9	9290521	19990412	6532453	20030311	Genetic programming problem solver with automatically defined stores loops and recursions
10	9336373	19990617	6424959	20020723	Method and apparatus for automatic synthesis, placement and routing of complex structures
11	9393863	19990910	6564194	20030513	Method and apparatus for automatic synthesis controllers
12	10355443	20030130	7117186	20061003	Method and apparatus for automatic synthesis of controllers

As mentioned earlier, aligning Mr. Koza's patents as listed in Table 2 provides little insight into the evolution of his EC research. While a subjective examination of their

document contents and prosecution histories may reveal how some patents are related to or directly derived from others, this alone does not objectively determine whether they follow a continuous line of research or originate from separate endeavors—let alone their relationship with other EC patents.

EC citation network

This study constructs a citation network using EC patent documents and their backward citations. A few key points about this construction are as follows:

1. Node Representation: The nodes in the network are identified by their patent application numbers. This arrangement aggregates citations for an invention's patent and its corresponding pre-grant publications (PGPubs), providing a more comprehensive view of its citation relationships (Kuan, Chen, & Huang, 2020).
2. Citation Ordering: All citations are filtered so that the cited patent documents are always those filed earlier than their citing counterparts. This prevents cycles in the network and ensures that knowledge flows consistently from earlier-filed patents to those filed later.
3. Network Closure: The network is closed, meaning that only patent documents classified as EC are included—both cited and citing—by filtering out those outside the EC patent dataset. While this restriction is not mandatory, it is applied for simplicity in analysis.
4. Removal of Anomalies: Loops and duplicate arcs are removed from the citation network. These anomalies result from the aggregation mentioned in (1). For example, a loop occurs when a patent self-cites its own PGPubs, while duplicate arcs appear when a later patent simultaneously cites an earlier patent and its PGPub.

After applying the aforementioned processing steps, the final EC citation network consists of 46,261 arcs and 19,836 nodes, representing approximately 54% of the 36,560 EC inventions. In other words, roughly half of the EC inventions lack mutual citations and are therefore not part of the citation network, suggesting potential imprecision in the AI Patent Dataset. However, this study verifies that all 12 of Mr. Koza's EC patents are included in the citation network.

The citation network is distributed across 1,178 components (i.e., isolated sub-networks). The largest component includes 16,757 nodes, accounting for approximately 85% of the total nodes, whereas all other components are significantly smaller (the second-largest component contains only 27 nodes). The MPs produced in subsequent analyses will be derived entirely from this largest component, as MPs do not extend across disconnected components.

A macro-view based on a researcher's or scientist' entire set of works

To derive Mr. Koza's research trajectory, this study assigns SPNP weights to the arcs of the citation network (Kuan, 2020). Subsequently, the 12 nodes, each corresponding to one of Mr. Koza's patents listed in Table 2, are gathered into a Pajek cluster, and the GKN and LKN searches are applied to generate Mr. Koza's individual-level GKN and LKN MPs. For simplicity, the LKN search is conducted

with a tolerance value of zero, meaning that only arcs with the topmost weight are traced.

The resulting GKN MPs include 54 nodes, while the LKN MPs include 69 nodes. Although the two sets of MPs differ—each containing some nodes absent from the other—both reflect a common theme in the evolution of Mr. Koza's research, as their interconnected structures share a consistent framework (as described below). Additionally, 50 out of the 54 nodes in the GKN MPs are also present in the LKN MPs. Therefore, for brevity, only the GKN MPs are presented in Figure 2.

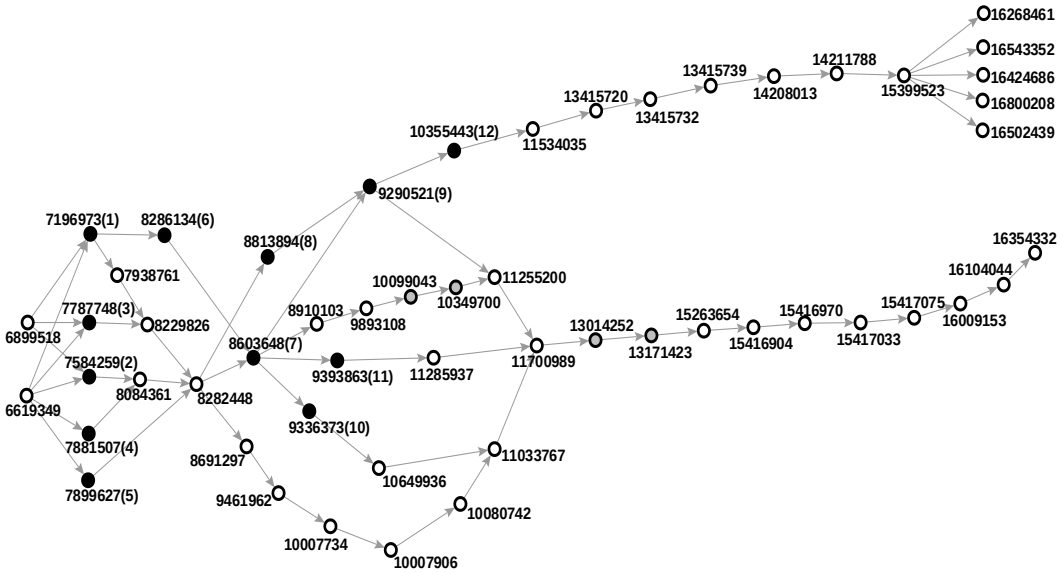


Figure 2. Mr. Koza's individual-level MPs by GKN search.

In Figure 2, the nodes are labeled with their corresponding patent application numbers. The black nodes represent Mr. Koza's 12 patents, with their sequence numbers from Table 2 in parentheses attached to their labels. The four grey nodes denote those that are not present in the LKN MPs.

At first glance, Figure 2 reveals that, as an early pioneer in the EC field, Mr. Koza's works are concentrated in the early (or left) half of the trajectory. All of his works can be traced back to a common origin. Then, Mr. Koza's works initiate two distinct strands of subsequent development in the later (or right) half of the trajectory.

As mentioned earlier, the LKN MPs reveal an identical framework to that depicted in Figure 2, except that they include an additional source, an additional sink, and several extra nodes and branches in the denser left portion of the trajectory.

A closer examination of the patents in Figure 2 reveals that the early half of Mr. Koza's individual-level MPs follows a development trajectory centered on the evolution of computer programs based on genetic algorithms. Interestingly, in the later half, the trajectory transitions toward neural network-based product design and the training of machine learning models.

The common origin of all 12 of Mr. Koza's patents involves two prior patents, both of which are based on genetic algorithms:

1. Application No. 6619349 (corresponding to Patent No. 4697242) lists John Holland as one of the inventors, who is recognized as the father of genetic algorithms (Bäck, Hammel, & Schwefel, 1997). This patent describes an adaptive computing system consisting of a population of classifiers. The system employs a genetic algorithm to generate new classifiers, replacing less effective ones and enabling continuous learning and improvement.
2. Application No. 6899518 (corresponding to Patent No. 4821333) describes a method for image recognition, particularly focusing on applying mutation and crossover mechanisms to evolve sets of structuring elements within an image. The goal is to identify an optimal set of structuring elements that can effectively distinguish between image categories.

For brevity, this study examines four patents, selecting two from each strand of subsequent developments. In the lower right part of Figure 2, the knowledge flow from Mr. Koza's genetic programming work shifts into the training of machine learning models:

1. Application No. 15263654 (corresponding to Patent No. 10387801) focuses on training and assessing a machine-learned model to refine a large collection of documents (e.g., web pages from a search engine) into a shorter ordered list (akin to a partial order). The ranking is derived from multiple parameters that reflect relevance, similar to fitness values in evolutionary algorithms.
2. Application No. 16354332 (corresponding to Patent No. 11494691) also focuses on training and assessing a machine learning model but specifically optimizes the training process. This more advanced patent introduces a technique that utilizes the idle time while the machine learning model awaits actual outcomes from its previous action. During this waiting period, the system generates a set of predicted outcomes and uses at least a subset of them to train the model, producing multiple candidate models—thereby accelerating the training process.

In the upper right part of Figure 2, the knowledge flow from Mr. Koza's work shifts separately into the domain of product design utilizing neural networks:

1. Application No. 11534035 (corresponding to Patent No. 8423323) discloses a system and method for designing new products. A mapping relationship between consumer preferences and product attributes is modeled using neural networks. Interactive Evolutionary Computation (IEC) and genetic algorithms are integrated to optimize the model and search process, allowing designers to predict the acceptance of new products and identify highly desirable yet underrepresented areas in the market.
2. Application No. 15399523 (corresponding to Patent No. 10783429) integrates artificial neural networks and evolutionary computation to automate the analysis of large-scale user data and efficiently identify the most effective web design. At the core of the system is a neural network that maps user attributes to different dimensions and values of a web page. The neural network is represented as a

genome and optimized through evolutionary operations such as initialization, testing, competition, and reproduction.

Despite the abridged description above, one can still discern a lineage of evolving ideas through the patents preceding and succeeding Mr. Koza's patents.

A micro-view based on a single work from a researcher or scientist

The previous section provides a macro-level perspective on Mr. Koza's individual-level MPs, demonstrating the usefulness of these MPs in identifying both the most influential sources contributing to his research and the most prominent subsequent developments arising from it as a whole.

However, this macro-view has limitations, as it does not explicitly clarify how Mr. Koza's specific patents are related to one another, nor how they connect to the identified sources and subsequent developments. For example, considering the patent Application No. 9290521, the macro-view alone does not help analysts determine whether it is more closely related to the upper strand of development, as suggested in Figure 2.

Additionally, Figure 2 shows that five of Koza's patents (numbered 1 to 5) appear in parallel immediately after their two common prior sources. However, the macro-view again fails to inform analysts whether they are equally related to Application No. 9290521. In fact, as will be demonstrated later, Figure 2 may even be somewhat misleading in answering these questions.

To overcome the shortcomings of the macro-view, this study proposes a micro-level perspective by conducting a GKN or LKN search on specific nodes representing the patents of interest, rather than designating all of Mr. Koza's patents as key nodes. To demonstrate the usefulness of this micro-view in supplementing the macro-view, this study performs a GKN and LKN search on a single key node, corresponding to Application No. 9290521.

Again, for brevity, only the resulting GKN key-node MPs are presented in Figure 3, as the differences between them and the LKN key-node MPs are minor. Similarly, in Figure 3, the black nodes represent Mr. Koza's patents (including 9290521), while the three gray nodes denote patents not present in the LKN key-node MPs.

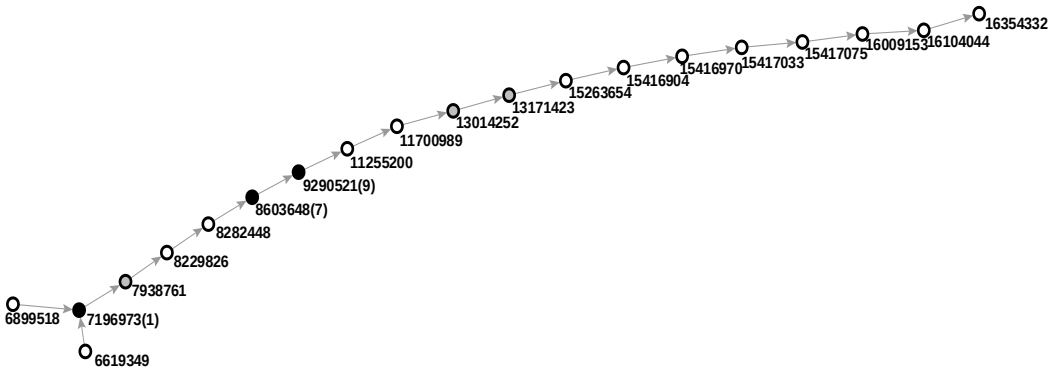


Figure 3. MPs from a single Mr. Koza's patent by GKN search.

Figure 3 reveals some unexpected observations. Firstly, two of Mr. Koza's patents, Application Nos. 8603648 and 7196973, are more significantly related to 9290521 than the others in terms of knowledge diffusion, as they are aligned along the same knowledge-diffusion lineage.

There are also two other patents adjacent to 9290521 besides 8603648 in Figure 2—Application Nos. 10355443 and 8813894. However, the key-node search selects 8603648, including it in 9290521's individual lineage.

As illustrated, this micro-view helps analysts differentiate the degree of relatedness between 9290521 and Mr. Koza's other patents, as well as understand its lineage. The same approach can be individually applied to each of Mr. Koza's patents.

Second, while 9290521 is structurally closer to the strand of subsequent development related to product design utilizing neural networks, its key-node MPs reveals that it is more closely aligned, in terms of knowledge diffusion, with the strand of subsequent development concerning the training of machine learning models.

Conclusion

This study contributes to the understanding of MPA by:

1. Promoting awareness of a previously overlooked path search algorithm in MPA, termed key-node search, which derives MPs extending both backward and forward from one or more key nodes.
2. Demonstrating the application of key-node search to capture researchers' or scientists' individual-level MPs, reflecting their research trajectories within the broader context of the field to which they belong.

Using a case study, this study demonstrates both a macro-view and a micro-view of a researcher's or scientist's individual-level MPs. The macro-view designates all of the researcher's or scientist's works as key nodes, helping to identify both the most influential sources contributing to the researcher's or scientist's research and the most prominent subsequent developments arising from it as a whole.

The micro-view, on the other hand, designates one or a few of the researcher's or scientist's works as key nodes. This supplements the macro-view by differentiating the degree of relatedness between these works and the researcher's or scientist's other works. Additionally, the micro-view provides insights into the relationship between these specific works and the most prominent subsequent developments uncovered in the macro-view.

While the individual-level MPs uncovered in the case study appear reasonable, the greatest challenge to the proposed approach lies in verifying how accurately these MPs reflect a researcher's or scientist's research evolution and how trustworthy the identified contributing sources and subsequent developments are.

Currently, analysts can only rely on subjective evaluation, experts' domain knowledge, or existing review articles and industry reports, if available. The issue of representativeness remains unresolved. However, this challenge is not unique to this study—it is a common limitation across all studies utilizing MPA.

There are several interesting extensions to this study. One such extension is that, instead of limiting the proposed approach to individual researchers or scientists, it could be applied to other types of "individuals," such as paper authors, research

institutes, or firms. The resulting individual-level MPs could then be interpreted as reflecting their research trajectories within the broader field.

Additionally, this study does not explore how one researcher's or scientist's individual-level MPs compare with those of another scientist or with the field-level MPs. Regarding the former, such an investigation could reveal how their research trajectories interact within the field—whether they run in parallel, converge, or diverge at certain points, among other patterns. Regarding the latter, examining interactions between individual-level and field-level MPs could uncover certain patterns, allowing researchers or scientists to be categorized based on their alignment with the field's mainstream development.

Open science practices

The data and software used in the case study are both publicly and freely available. The AI Patent Dataset can be downloaded from USPTO's website (https://www.uspto.gov/ip-policy/economic-research/research-datasets/artificial-intelligence-patent-dataset). The software Pajek can be downloaded from its official website (http://mrvar.fdv.uni-lj.si/pajek/).

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 113L900202), which is under The Featured Areas Research Center Program by the Ministry of Education (MOE) in Taiwan, and by the National Science and Technology Council (NTSC), Taiwan, under Grant No. 112-2221-E-011-116-MY2.

References

Back, T., Hammel, U., & Schwefel, H. P. (1997). Evolutionary computation: Comments on the history and current state. *IEEE transactions on Evolutionary Computation, 1*(1), 3-17.

Batagelj, V., & Mrvar, A. (1998). Pajek - Program for large network analysis. *Connections, 21*(2), 47-57.

Chen, P. H., & Cho, R. L. T. (2023). The Technological Trajectory of LEO Satellites: Perspectives from Main Path Analysis. *IEEE Transactions on Engineering Management*.

Cho, R. L. T., Liu, J. S., & Ho, M. H. C. (2021). The development of autonomous driving technology: perspectives from patent citation analysis. *Transport Reviews, 41*(5), 685-711.

De Nooy, W., Mrvar, A., & Batagelj, V. (2018). *Exploratory social network analysis with Pajek: Revised and expanded edition for updated software* (Vol. 46). Cambridge University Press.

Feng, S., & Magee, C. L. (2020). Technological development of key domains in electric vehicles: Improvement rates, technology trajectories and key assignees. *Applied Energy, 260*, 114264.

Giczy, A. V., Pairolero, N. A., & Toole, A. A. (2022). Identifying artificial intelligence (AI) invention: A novel AI patent dataset. *The Journal of Technology Transfer, 47*(2), 476-505.

Han, F., Yoon, S., Raghavan, N., Yang, B., & Park, H. (2024). Technological trajectory in fuel cell technologies: A patent-based main path analysis. *International Journal of Hydrogen Energy, 50*, 1347-1361.

Ho, J. C., Saw, E. C., Lu, L. Y., & Liu, J. S. (2014). Technological barriers and research trends in fuel cell technologies: A citation network analysis. *Technological Forecasting and Social Change, 82*, 66-79.

Hummon, N. P., & Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks, 11*(1), 39-63.

Kuan, C. H. (2020). Regarding weight assignment algorithms of main path analysis and the conversion of arc weights to node weights. *Scientometrics, 124*(1), 775-782.

Kuan, C. H. (2024). Integrating prior field knowledge as key documents with main path analysis utilizing key-node path search. *Journal of Informetrics, 18*(3), 101569.

Kuan, C. H., Chen, D. Z., & Huang, M. H. (2020). The overlooked citations: Investigating the impact of ignoring citations to published patent applications. *Journal of Informetrics, 14*(1), 100997.

Kuan, C. H., & Liao, S. Y. (2024). Assessing main paths by uncovering their coverage with key-node path search. *Scientometrics, 129*(11), 6629-6657.

Liu, J. S., & Lu, L. Y. (2012). An integrated approach for main path analysis: Development of the Hirsch index as an example. *Journal of the American Society for Information Science and Technology, 63*(3), 528-542.

Liu, J. S., Lu, L. Y., & Ho, M. H. C. (2019). A few notes on main path analysis. *Scientometrics, 119*, 379-391

Mitchell, M., & Taylor, C. E. (1999). Evolutionary computation: an overview. *Annual Review of Ecology and Systematics, 30*(1), 593-616.

Park, H., & Magee, C. L. (2017). Tracing technological development trajectories: A genetic knowledge persistence-based main path approach. *PloS One, 12*(1).

Park, H., & Magee, C. L. (2019). Quantitative identification of technological discontinuities. *IEEE Access, 7*, 8135-8150.

Su, F. P., Chen, S. J., Chang, Y. H., & Lai, K. K. (2019). Construct a three-stage analysis model of integrated main path analysis and patent family-exploring the development of blockchain. In *Proceedings of the 2019 3rd International Conference on Software and e-Business* (pp. 151-156).

Verspagen, B. (2007). Mapping technological trajectories as patent citation networks: A study on the history of fuel cell research. *Advances in complex systems, 10*(01), 93-115.

Watanabe, I., & Takagi, S. (2021). Technological trajectory analysis of patent citation networks: examining the technological evolution of computer graphic processing systems. *The Review of Socionetwork Strategies, 15*, 1-25.

https://doi.org/10.51408/issi2025_080

Exploring the Application of Open Peer Review in Academic Evaluation: An Analysis of H1 Connect Recommended Papers

Liu Xiaojuan[1], Xiang Nannan[2], Yu Yao[3]

[1] lxj_2007@bnu.edu.cn, [2] bnu_xnn@163.com, [3] yuyao990824@163.com

School of Government, Beijing Normal University, No. 19, Xinjiekouwai Street, Haidian Beijing (China)

Abstract

The development of open peer review has provided a new perspective on academic evaluation. By exploring the relationship between peer review indicators and impact indicators including Citation and AAS, as well as delving into the value of papers from the perspective of peer reviewers, this research offers insights to improve academic evaluation systems. The study focuses on papers about three topics: *Cardiovascular Diseases*, *Respiratory Tract Diseases*, and *Neoplasms*. It utilizes open peer reviews from H1 Connect, analyzing them from two dimensions: review indicators and scientific research contributions. Regarding review indicators, attention is paid to the RNumber and the RStar. The analysis of contributions is based on the Becker Medical Library's research evaluation model, which is used to design a classification system for contribution types. This study employs the "GPT-4O-mini" model to extract sentences describing scientific research contributions from peer review texts, and then categorizes them according to the designed classification system. The findings reveal that, in terms of review indicators, there are significant differences across topics, with a notable positive correlation existing between the RNumber, RStar, Citation, and AAS. In terms of scientific research contributions, these contributions are primarily concentrated in the dimensions of Knowledge Advancement and Clinical Implementation, with slight differences in contribution types among the topics. Contributions regarding clinical trial outcomes and healthcare services are more prominent in *Cardiovascular Diseases*, while theoretical contributions are more apparent in *Respiratory Tract Diseases*. Regarding contribution co-occurrence, Knowledge Advancement and Economic and Community Benefits contributions often do not occur simultaneously. Papers that contribute to the discovery of new ideas, data methods, or clinical management and treatment are more likely to exhibit multiple types of contributions. Contributions to public health policies often appear separately. Generally, papers tend to focus on making significant contributions in one specific area, with the occurrence of multiple types of contributions being relatively rare. Academic evaluation should effectively integrate peer review with impact indicators, while deeply exploring the scientific research contributions of papers. It is crucial to consider both the diversity of contributions and the thematic differences to build a more comprehensive, scientific, and effective academic evaluation system.

Introduction

Peer review is a key mechanism to ensure the quality of publications, maintain academic integrity, and promote scholarly communication. The peer review process is intended to help improve research reporting and weed out work that does not meet the research community's standards for research production (Wolfram et al., 2020). It relies on the expertise and judgment of experts in the field to evaluate academic papers, project proposals, or research achievements to determine whether they fulfill the criteria for publication or funding. However, the traditional peer review process is often regarded as a closed and opaque "black box operation". Its information such as the decision basis, review texts, and reviewer identity is often not disclosed, which not only limits the transparency of the research process but also may lead to bias (Demarest et al., 2014; Fox & Paine, 2019), unfairness (Bravo et al., 2018) and inefficiency.

Open Science came into being to improve the transparency, fairness, and efficiency of scientific research and evaluation. It has become an important concept to promote the sustainable development of scientific research. Open Science advocates the openness and transparency of all facets of scientific research, and open peer review (OPR) is the last frontier of Open Science that has yet to achieve widespread adoption (Wang et al., 2016), has gradually become one of the means to overcome some limitations of traditional peer review.

Through the efforts of relevant institutions to enhance the transparency of the academic publishing process and oversee the peer review work, the credibility of peer review can be improved. This helps reduce unjust, unprofessional, and unnecessary evaluations of papers, thereby advancing the goals of follow-up reviews, peer review accountability, and review quality supervision (Wang, 2023). An increasing number of journals and conferences have started to implement the open peer review mechanism in recent years, and open peer review platforms such as H1 Connect, Publons, and Pubpeer have also emerged. These platforms significantly lessen the difficulty of obtaining peer review data and further enrich the types of peer review data, including review texts, review scores, review numbers, etc. Open peer review data are the tangible representation of expert opinions, with greater professionalism, transparency, and credibility than traditional citation data and altmetrics data. It also has rich value, offering a foundation for investigating the behavior of peer review, identifying the traits of expert reviews, and exposing the peer review process's working mechanism.

Focused on the open peer reviews from H1 Connect, this study analyzes the reviews from two dimensions: the numerical characteristics and the scientific research contributions. Additionally, impact indicators, including Citation and Altmetric Attention Score (hereinafter abbreviated as AAS), are incorporated to explore the relationship with peer review indicators. By integrating these dimensions and impact indicators, our goal is to delve deeper into the realm of peer reviews and investigate their potential value in academic evaluation. To be more specific, this study seeks to answer the following questions: What distribution characteristics can be observed in open peer review indicators? Are there differences across research topics? What is the relationship between peer review indicators, Citation and AAS? Do papers that receive higher recognition from peer reviewers tend to achieve higher impact? What research contributions are embedded in open peer review texts, and how are these contributions distributed and co-occurring?

Literature review

With the growing momentum of the open science movement, an increasing number of journals and publishers are joining the ranks of those sharing peer review data. Meanwhile, numerous open peer review platforms have emerged, laying a practical groundwork for peer reviews exploration. The development of technologies such as natural language processing and sentiment analysis provides technical support for the implementation of peer review mining. In addition to the review comments in the form of text (hereinafter abbreviated as "peer review texts"), there are also various forms such as review scores, review numbers, review stars, and review labels. Many scholars have carried out analysis and utilization research on different types of peer reviews.

Open peer review, Citation and AAS

Some studies have explored the effect of open peer review on citation and AAS. Zong et al. (2020), using PeerJ as an example, found that articles with open peer review history could be expected to have significantly higher citations than those with a traditional review pattern, but there would be variations among disciplines. However, some investigations reach different conclusions. According to Ni et al. (2021), there is no evidence of a citation advantage for the papers disclosing their peer review documents by taking *Nature Communications* as an example. Articles subjected to OPR have no obvious advantage in citation but a notably higher score in altmetrics (Cheng et al., 2024). Xie et al. (2024) revealed that different types of

papers have significant differences in review scores and citations, and there is a positive correlation between review scores and citations.

Sentiment analysis of peer reviews

Peer review texts often contain rich sentimental information, reflecting the reviewers' overall attitude toward the research presented in the paper. Therefore, sentiment analysis is widely employed in peer review text mining, and most studies aim to classify the sentiment polarity of peer review texts. Wang et al. (2018) introduced sentiment analysis into peer review texts analysis for the first time. By using automatic identification, they detected sentence fragments with positive or negative connotations. These fragments, representing sentiment polarity, were then used to predict the final score of a paper. Based on the sentiment information in the authors' comments and the content of the peer review texts, Ghosal et al. (2020) developed the DeepSentiPeer model to forecast the overall recommendation score and ultimate decision of the work. Bravo et al. (2019) examined whether the language style of the reviewers changed after the journal opened the peer review report, using continuous numerical values to represent the sentiment polarity and subjectivity of the review texts. Lin et al. (2021) employed the sentiment analysis model to mine the sentiment polarity of open review texts. They used the titles, abstract, Twitter comments, and peer review texts as input to the model, with the average review scores as the actual score. The evaluation of the paper was based on the sentiment polarity of the review texts. Some scholars have further combined the sentiment polarity of peer review texts with citations. Zong et al. (2020) investigated the relationship between the sentiment polarity of peer review comments and citations using data from PubPeer, F1000, and ResearchGate. They discovered that in comparison to the comparable control pairings (articles without PPPRs), papers that obtained favorable post-publication peer reviews (PPPRs) had noticeably higher citations. However, the control group, which included papers with neutral or negative ratings and papers with both positive and negative reviews, did not differ significantly in citations.

Identification of elements in peer review text

Peer review texts on academic papers are typically long and structurally complex. Identifying the elements contained in them can help gain a deep understanding of the peer review mechanism and its value. At present, many studies have defined and identified the types of elements from different perspectives. Hua et al. (2019) divided the elements into evaluation, request, fact, reference, and quote. They then examined the effects of several models on element identification and found that the Bi-LSTM-

CRF model had the best effect. Fromm et al. (2021) separated the elements into non-arguments, supporting arguments, and opposing arguments, and tested the performance of the Bert model in the argument extraction task. They also pointed out that peer review texts differ from other types of subjective texts (such as legal documents and e-commerce reviews) in terms of length, tone, and wording. Chen et al. (2023) separated the elements into four categories, including overview, method, result, and highlights, using the step type definition in conjunction with the research corpus's content characteristics. They then evaluated the recognition efficacy of SVM, FastText, TextCNN, and BiLSTM models, concluding that the BiLSTM model performed the best. Ghosal et al. (2022), using the ICLR peer review dataset as an example, categorized peer review texts into four dimensions: the section of the paper that the review comments on (e.g., Introduction, Methodology, Data, Experiments), the aspect of the paper that the review addresses (e.g., Appropriateness, Originality or Novelty, Clarity), the purpose or the role of the review (e.g., Suggestion, Discussion, Question), and the significance of the review (e.g., Major Comment, Minor Comment, General Comment). Zhang et al. (2022), using 3329 comments from 690 papers published in the British Medical Journal (BMJ) as the research objects, analyzed the differences in the length distribution of reviewers' comments, the general distribution of words in comments and the position of reviewer comments. Wang et al. (2020) analyzed the review texts of papers published in journals such as *Cell* and *The Lancet* recommended by F1000Prime and found that the most frequently used words by experts included interesting, important, first, exceptional, etc.

In summary, existing research primarily focuses on the analysis of open peer review indicators, sentiment analysis, and element recognition based on peer review texts. While some studies examine the characteristics of open peer review data from various perspectives, most of them address only a limited number of indicators and rarely consider the inherent characteristics of the papers themselves. In this study, we take a more comprehensive approach by analyzing both the textual and numerical aspects of peer review data. Methodologically, most existing studies rely on machine learning and deep learning models to analyze the content of review texts. However, the generalizability, adaptability, and enhanced capabilities of large language models in feature extraction, semantic understanding, and multimodal learning provide models with significant advantages in identifying elements within peer review texts. In this study, we introduce large language models to extract research contributions. The aim is to comprehensively reveal the value of papers from the perspective of peer reviewers, enhance the understanding of post-publication open peer reviews,

and provide insights into the application of peer review in academic evaluation within the context of open science.

Data and method

Data

Among the many open peer review platforms, H1 Connect (formerly F1000, F1000 Prime and Faculty Opinions) has been the most authoritative representative in the global biomedical field in the past twenty years. It brings together nearly ten thousand top experts in the field, aiming to further recommend and evaluate papers that have been published after traditional peer review. Therefore, this study uses H1 Connect as the source of open peer review data.

Neoplasms, *Cardiovascular Diseases*, and *Respiratory Tract Diseases* are characterized by high morbidity and mortality, severely affecting human health. This study selected academic papers on these three topics for research. First, a search was conducted in the PubMed database using "Neoplasms," "Cardiovascular Diseases," and "Respiratory Tract Diseases" as MeSH Major Topics, with the time frame limited to January 2015 to December 2020, and the document types restricted to "Article" or "Review." A total of 1,496,535 papers were retrieved, of which 10,810 were recommended by H1 Connect, including 9,580 articles and 1,230 reviews. Among the recommended papers, there are 3,526 papers on *Cardiovascular Diseases* (hereinafter abbreviated as C), 2,488 papers on *Respiratory Tract Diseases* (hereinafter abbreviated as R), and 5,640 papers on *Neoplasms* (hereinafter abbreviated as N). It should be noted that some papers belong to multiple topics simultaneously. Next, we collected the Citation and altmetrics data for the recommended papers using their DOI from Web of Science and Altmetric.com. Finally, we used a self-written Python program to scrape open peer reviews on H1 Connect, obtaining a total of 12,203 reviews The final dataset collected includes the topic, paper title, publication year (hereinafter referred to as Year), Citation, AAS, type of document (hereinafter referred to as Type), review number (hereinafter abbreviated as RNumber), review star (hereinafter abbreviated as RStar), and review text. The distribution of papers by publication year is shown in Table 1.

Table 1. Publication year distribution of papers.

Year	*Paper count*
2015	2023
2016	1999
2017	1939
2018	1775
2019	1657
2020	1417
Whole	10810

Method

(1) Extraction of scientific research contribution sentences

The scientific research contribution refers to the ability of the current research to improve, perfect and apply existing knowledge, theories or practices (Luo et al., 2021), including new theories, new methods, new technologies, new outcomes. Analyzing and evaluating these contributions is a necessary step in evaluating the quality of the paper and promoting knowledge innovation and disciplinary progress. Previous studies analyzing the scientific research contributions of papers were mainly based on abstract or full-text datasets and relied mainly on the authors' descriptions, which introduces a certain degree of subjectivity. In contrast, the insights and evaluations in peer review texts come from authoritative experts, making them an important reference for uncovering the paper's scientific research contributions. Therefore, this study further explores the scientific research contributions of papers based on peer review texts.

Traditional deep learning models rely heavily on large-scale, high-quality annotated data. The powerful contextual understanding ability of large language models enables them to achieve excellent performance in downstream tasks with only a small number of examples or direct prompts, thereby shifting the paradigm of information extraction tasks from fine-tuning to zero-shot/few-shot (Shi et al., 2024). This study uses the "GPT-4O-mini" model to extract scientific research contribution sentences from peer review texts. Firstly, combining with the definition of scientific research contribution, this study argues that the scientific research contribution sentence in peer review texts should meet both of the following conditions: (1) The sentence must explicitly mention the study. (2) The sentence must express the experts' recognition of the study's value. This study designs the model prompt based on this, as shown in Figure 1. Secondly, a test sample of 1,000 review texts was

constructed and manually annotated according to the two conditions. The extraction performance of zero-shot, one-shot, and few-shot prompt strategies is tested respectively. Through experiments, it is found that in a few cases, the model's output might slightly change the original sentence. Therefore, further processing of the model's extraction results is necessary. By writing code to determine whether the extracted sentences are the original sentences of the review text, we match the sentences that do not meet the requirements to the original text based on cosine similarity. Next, the micro-average index is used to evaluate the extraction performance of different prompt strategies, and then the strategy with the best performance is selected to extract all review texts. Finally, the extraction results are manually verified.

You are a top scholar in the field of medicine. Below is a peer review text from an expert regarding a medical research paper. Please extract all the sentences from the expert that evaluate the value of the paper. Each extracted sentence must meet both of the following criteria:

1. The sentence must explicitly mention the study (e.g., "this study," "this article," "this paper," or specific descriptions of the study's content).
2. The sentence must express recognition of the study's value (e.g., "help," "valuable," "important," "novel," or "new method"). Avoid sentences that merely describe the study's content without including a value judgment.

Output requirements:

1. Only output the original extracted sentence(s) exactly as written, with no changes.
2. If multiple sentences are extracted, separate each sentence with a "\n\n".
3. If no relevant content is found, output "None.".

Here is an example:

Input: peer review text.

Output: sentence1. \n\n sentence2……

Input: ……

Figure 1. Model prompt.

The extraction performance of different prompt strategies is shown in Table 2. It can be found that the optimal F_1 value of the zero-shot strategy can reach 74.42%. Therefore, the zero-shot prompt strategy is used to extract all peer review texts. The scientific research contribution sentences have been extracted, totaling 7,290(including 279 non-original sentences, accounting for only 3.83%). After manual verification and filtering, 5021 sentences remain, involving a total of 3207 papers.

Table 2. Performance of different prompt strategies.

Prompt strategy	*P*	*R*	F_1
0-shot	71.14%	78.01%	74.42%
1-shot	69.08%	75.31%	72.06%
few-shot	59.94%	80.08%	68.56%

(2) Classification of scientific research contributions

The Becker Medical Library Research Evaluation Model, designed by Washington University School of Medicine, aims to go beyond traditional citation analysis indicators to comprehensively assess the value and impact of medical research. The model tracks research output, dissemination, and transformation, providing a comprehensive evaluation of biomedical research across five dimensions: advancement of knowledge, clinical implementation, legislation and policy, economic benefit, and community benefit. This study refers to the model and combines the actual characteristics of the extracted sentences to divide scientific research contributions into nine types from three dimensions. Relevant explanations and examples can be found in Table 3. Manual annotation is conducted based on this categorization system.

Table 3. Classification and explanation of the types of scientific research contributions.

Contribution Type	*Contribution Subtype*	*Explanation*	*Example*
1 Knowledge Advancement: Research outcomes contribute to the expansion and promotion of the knowledge system	1.1 Concepts & Theories	Initiating new research directions; proposing new theoretical frameworks, concepts, or hypotheses.	This study **creates a new paradigm** in critical care medicine.
	1.2 Insights & Findings	Formulating new insights, findings, conclusions, or confirmations during the research process.	These observations **add significant new insights** to our understanding.....

	1.3 Data & Methods	Constructing meaningful datasets; proposing or improving new methods, strategies, or pathways to research questions.	This paper **reports on a new methodology** .....
2 Clinical Implementation: Research outcomes contribute to the improvement of clinical practice	2.1 Medical Products	Research outcomes aid in the selection and development of medical products, such as pharmaceuticals, biomaterials, and medical devices.	Such genome-wide systematic and unbiased strategies could **help in developing a wide range of drugs**.....
	2.2 Clinical Management and Treatment	Research outcomes contribute to clinical decision-making, optimizing clinical management, or enhancing clinical treatment plans.	The data therefore **open new therapeutic avenues.**
	2.3 Clinical Trial Outcomes	Clinical trials have achieved valuable outcomes.	**The WINTHER clinical trial provides a glimpse of the value of** .....
3 Economic and Community Benefits: Research outcomes can enhance economic benefits or improve community welfare	3.1 Healthcare Services	Improving health conditions; enhancing health literacy; reducing service costs.	This may **help to lower resource use, costs, and enhance quality and value of care.**
	3.2 Morbidity & Mortality	Alleviating the disease burden; decreasing morbidity and mortality rates;	This review has **important implications for prevention** of VTE as a major cause of

	increasing survival rates.	maternal **mortality and morbidity**.
3.3 Public Health Policy	Providing a scientific basis for the formulation of public health policies, guidelines and related measures.	This study **justifies the policy**.....

Results

Impact indicators of recommended papers

Citation analysis evaluates academic impact within a specific discipline; altmetrics emphasizes social impact on the public, and peer review provides an in-depth evaluation of a paper's content from an expert perspective. To analyze the characteristics of recommended papers from multiple perspectives and provide a reference for subsequent comparative analysis of peer review comments, this study first analyzes two commonly used impact indicators, Citation and AAS, to explore the impact of the recommended papers.

(1) Citation. The citation of recommended papers (Table 4, Figure 2) is highly dispersed, with a large span, ranging from a minimum of 0 to a maximum of 21,917, and an average of 203.95. Among them, papers on *Respiratory Tract Diseases* have a higher average Citation (243.94) and are the most dispersed, while the Citation of papers on *Cardiovascular Diseases* is generally concentrated at a lower level. Among these recommended papers, papers on *Respiratory Tract Diseases* have a higher effect on the academic community.

Table 4. Distribution of Citation on different topics.

	Mean	*Min*	*Max*	*SD*
Whole Data	203.95	0	21917	557.9242
Cardiovascular Diseases	163.96	0	5885	396.77
Respiratory Tract Diseases	243.94	0	21917	836.66
Neoplasms	216.57	0	9728	488.99

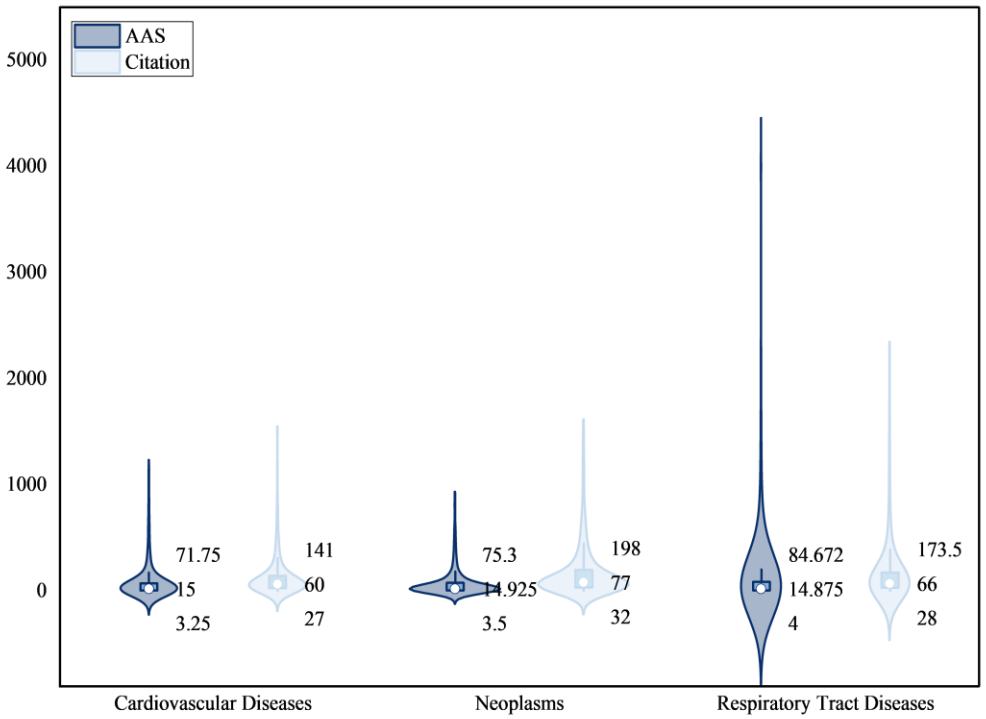


Figure 2. Distribution of impact indicators on different topics.

(2) AAS. The distribution of AAS is similar to that of Citation, with the data being highly dispersed and spanning a large range, from 1 to 32,243.46 (Table 5). The average values of AAS for each topic are significantly different. The papers on *Respiratory Tract Diseases* have a comparatively higher AAS, with an average value of two to four times that of other topics. While the AAS of papers on *Neoplasms* are concentrated at lower levels and have a lower degree of dispersion. It is evident that there are differences in the level of public attention towards papers on different topics. Papers belonging to *Respiratory Tract Diseases* generally have a higher and more scattered social impact, while papers on *Neoplasms* show more consistent levels of social attention.

Table 5. Distribution of AAS on each topic.

	Mean	*Min*	*Max*	*SD*
Whole Data	152.91	1	32243.46	782.56
Cardiovascular Diseases	120.60	1	12737.04	420.38
Respiratory Tract Diseases	341.31	1	32243.46	1531.80
Neoplasms	89.97	1	7380.74	250.70

Preliminary analysis reveals differences in the impact of papers across the three topics. To further examine these differences, this study performs differential tests on Citation and AAS. Due to the non-normal distribution of the data, the Kruskal-Wallis H test is conducted to analyze the data differences both among the three topics and between pairs of topics, as shown in Table 6. There is a statistically significant difference in Citation among the three topics ($H=70.682$, $p<0.001$), and a significant difference in AAS among the three topics ($H=10.820$, $p<0.01$). An analysis of pairwise topic differences is conducted, with each row in the table testing the null hypothesis that "the distributions of Topic 1 and Topic 2 are the same." The significance values have been adjusted by Bonferroni correction for multiple comparisons. Regarding Citation, significant differences in data distributions were observed between all pairs of the three topics. However, for AAS, the differences between topics varied. *Neoplasms* showed significant differences in data distributions compared to the other two topics, while the AAS data distributions for *Cardiovascular Diseases* and *Respiratory Tract Diseases* were nearly the same.

Table 6. Differential test of Citation and AAS across topics.

	H	*P_value*	*Group*	*P_value*
Citation	70.682	0.000***	C-R	0.003**
			C-N	0.000***
			R-N	0.000***
AAS	10.820	0.004**	C-R	1.000
			C-N	0.010**
			R-N	0.008**

* p<=0.05 ** p<=0.01 *** p<=0.001

Peer review indicators of recommended papers

(1) RNumber. The number of reviews can reflect the degree of attention paid by the experts to the paper. The average of RNumber is 1.13, with the majority (90.63%) of papers recommended only once by experts, and a very small proportion (0.22%) receiving 5 or more recommendations. The highest RNumber obtained by a paper is 11. The pairwise distribution of the RNumber and the RStar is shown in the center scatter plot of Figure 3, where it is evident that there is a linearly positive correlation between the RStar and the RNumber of the paper. Papers with a higher RNumber tend to receive higher RStar. The distribution of RNumber and RStar for papers

under various topics is displayed in the box plots on the top and right sides, respectively. It demonstrates that the publications on *Respiratory Tract Diseases* have been recommended comparatively more frequently and are distributed more widely.

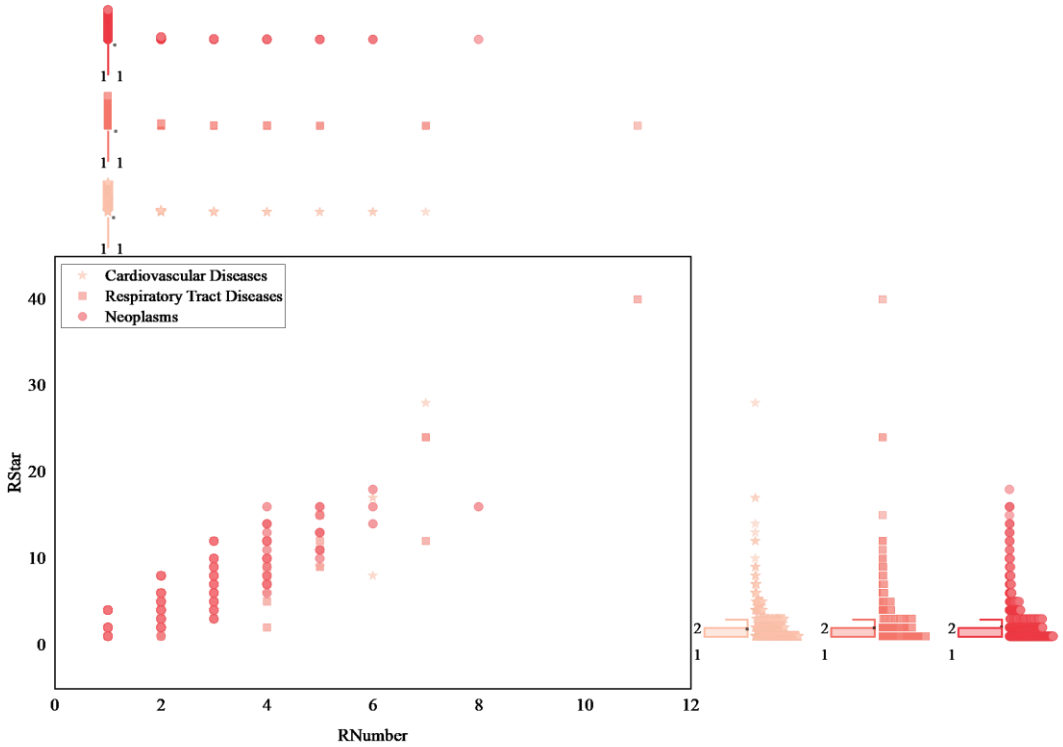


Figure 3. RNumber and RStar of papers on different topics.

(2) RStar. RStar can reflect the experts' recognition of the content and value of the paper. The RStar has a wide range, with a minimum of 1 and a maximum of 40. The mean and standard deviation of the RStar value are 1.96 and 1.64, respectively. RStar is typically low and concentrated. Only 0.46% of papers have an RStar of greater than ten, while the majority of papers (83.65%) are concentrated between one and two. Nearly half (49.70%) of the papers have an RStar of one. The papers on *Respiratory Tract Diseases* have a comparatively high RStar. Three of the four papers with RStar more than twenty are related to *Respiratory Tract Diseases*, while one belongs to *Cardiovascular Diseases*. On average, papers on *Neoplasms* received a higher average of 2.04, and the span of RStar obtained was also the smallest (1~18). In terms of dispersion, the RStar of papers on *Cardiovascular Diseases* is the most concentrated, while those on *Respiratory Tract Diseases* are the most dispersed.

To further investigate whether there are differences in distributions of peer review indicators among papers with different topics, we conducted difference tests on RNumber and RStar, respectively. Given that the tested data exhibited a non-normal distribution, non-parametric tests were employed. Specifically, this study utilized the Kruskal-Wallis H test to analyze the differences in data among three topics and between each pair of topics. The results are presented in Table 7. The results indicated the presence of statistically significant differences in RNumber among the three topics ($H=10.860$, $p<0.001$), as well as marked differences in RStar across these topics ($H=38.837$, $p<0.001$). These results suggest that experts taking part in open peer review have varying levels of attention and recognition towards papers of distinct topics. Pairwise comparisons of topic differences were conducted. The null hypothesis that "the distributions of Topic 1 and Topic 2 are the same" was tested for each row in the table. The Bonferroni correction method was used to modify the significance values for multiple tests. In terms of the RNumber, *Cardiovascular Diseases* shows significant differences from the other two topics, while *Respiratory Tract Diseases* and *Neoplasms* are nearly the same. In terms of the RStar, *Neoplasms* is significantly different from the other two topics, while there is no significant difference between *Cardiovascular Diseases* and *Respiratory Tract Diseases*.

Table 7. Differential test of RNumber and RStar across topics.

	H	*P_value*	*Group*	*P_value*
RNumber	10.860	0.004***	C-R	0.030*
			C-N	0.006**
			R-N	0.884
RStar	38.837	0.000***	C-R	0.132
			C-N	0.000***
			R-N	0.003**

* p<=0.05 ** p<=0.01 *** p<=0.001

Correlation test between peer review indicators and impact indicators of recommended papers

After conducting the Kolmogorov-Smirnov (K-S) test, it was determined that the RNumber, RStar, Citation, and AAS did not follow a normal distribution. Therefore, Spearman's correlation analysis was employed to assess the correlations among these indicators, with Spearman's rank correlation coefficient serving as a measure of the strength of these relationships. The results of the correlation test are presented in

Figure 6. The upper right triangular area indicates the significance levels of the correlations, with the shape and color of the ellipses representing the positive or negative nature of the correlations. Positive correlations are depicted as upward-facing ellipses, where a darker color signifies a stronger correlation. The numerical values in the lower left triangular area represent the correlation coefficients, with values closer to 1 indicating stronger positive correlations. It is observed that there is a significant positive correlation between the open peer review indicators of papers and impact indicators. Within each group of indicators, namely between RNumber and RStar, as well as between Citation and AAS, there are also significant positive correlations. Among them, the positive correlation between RNumber and RStar, and between Citation and AAS is higher (the two correlation coefficients are 0.68 and 0.48, respectively). This means that papers with more recommendations would be given higher review stars, and similarly, papers with higher citations would be given higher AAS.

RStar has a stronger positive effect on impact indicators than RNumber. The number of reviews positively affects both the Citation and AAS of a paper to a similar extent. The more reviews a paper receives, the wider its dissemination in academia and society, and the greater its impact. There is a strong positive correlation between RStar, Citation, and AAS, with RStar exerting a somewhat stronger positive effect on AAS. This suggests that papers that receive more positive reviews from experts will have higher citations and AAS.

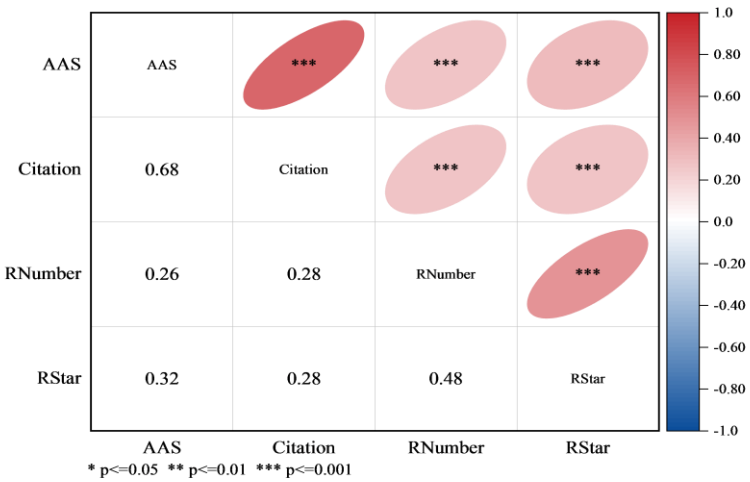


Figure 4. Correlation Test.

Spearman's correlation test was further performed on RNumber, RStar, Citation, and AAS under each topic, and the results are shown in Table 8. The correlation test findings of these variables for each topic show substantial positive correlations, which are basically consistent with the overall data. Notably, the strongest link is seen between Citation and AAS. The degree of correlation among different topics across various indicators varies. Except for a somewhat lower correlation between AAS and RStar compared with the situation in *Neoplasms*, *Respiratory Tract Diseases* shows stronger relationships among all indicators than the other two topics. *Cardiovascular Diseases* has the poorest positive correlations among the indicators.

Table 8. Correlation test for different topics.

	Cardiovascular Diseases				*Respiratory Tract Diseases*			
	AAS	*Citation*	*RNumber*	*RStar*	*AAS*	*Citation*	*RNumber*	*RStar*
AAS	1.000				1.000			
Citation	.661**	1.000			.703**	1.000		
RNumber	.246**	.265**	1.000		.275**	.303**	1.000	
RStar	.223**	.192**	.453**	1.000	.323**	.328**	.501**	1.000

	Neoplasms			
	AAS	*Citation*	*RNumber*	*RStar*
AAS	1.000			
Citation	.699**	1.000		
RNumber	.272**	.270**	1.000	
RStar	.367**	.295**	.491**	1.000

Scientific research contributions of recommended papers

(1) Distribution of scientific research contributions

The review text can reflect various contributions of the paper in different aspects (Qin, 2020). Figure 4 shows the distribution of the nine contribution types in the three dimensions involved in the review. The paper's contributions are more prominent in the areas of knowledge advancement, followed by clinical implementation, with relatively less emphasis on economic and community benefits. Among these, the reviewers focus more on the insights and findings of the paper, its value for clinical management and treatment, and the data and methods used in the paper. This aligns with the findings of previous research. Some studies on the reviews of academic papers in different fields have found that research methods, as an important part of the paper, are the focus of the reviewers (Han et al., 2022; Qin, 2020).

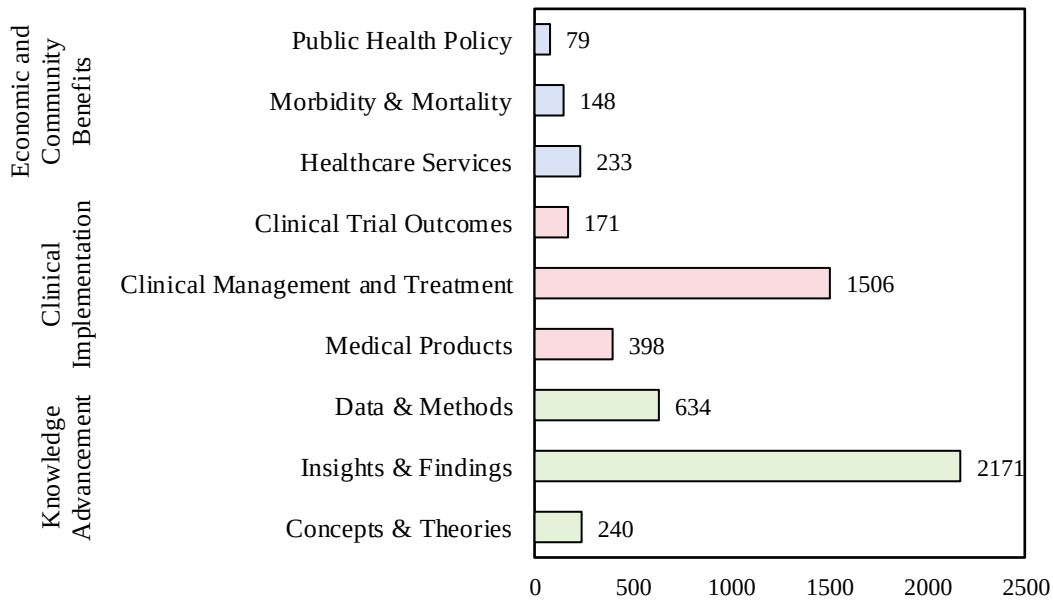


Figure 5. Distribution of types of scientific research contributions.

To better understand how scientific research contributions vary across papers on different topics, further exploration of their distribution is necessary (Figure 5). Analogous to the overall situation, it is observed that contributions in terms of insight discovery, clinical management and treatment, as well as data and methods dominate across all three topics. Slight variations exist among these topics. Specifically,

contributions related to clinical trial outcomes and healthcare services are more pronounced in *Cardiovascular Diseases* compared with the other two topics, whereas conceptual and theoretical contributions are more evident in *Respiratory Tract Diseases*.

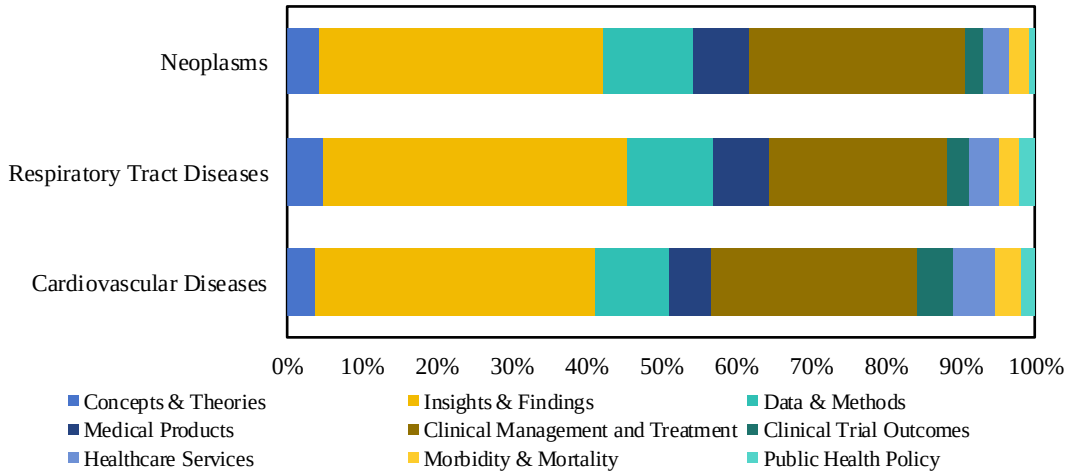


Figure 6. Distribution of the types of scientific research contributions of papers on different topics.

(2) Co-occurrence analysis between different types of scientific research contributions

Based on the overall distribution analysis of contribution types, to analyze the co-occurrence between different types of contributions helps gain a deeper understanding of the relationships or the influences between various contribution types. There are 2,145 papers demonstrating contributions in Knowledge Advancement, 1,669 papers exhibit contributions related to Clinical Implementation, and 413 papers present contributions in terms of Economic and Community Benefits. Notably, contributions of Knowledge Advancement and Clinical Implementation types tend to coexist more frequently, with 693 papers exhibiting both types of contributions. Following this, the coexistence of Clinical Implementation and Economic and Community Benefits contributions is observed in 241 papers. The coexistence of Knowledge Advancement and Economic and Community Benefits contributions is the least prevalent, occurring in 183 papers. Additionally, 97 papers exhibit contributions across all three types.

It is demonstrated that breakthroughs in basic research often propel advancements in clinical practice. This close connection may stem from the trend in modern medical research. Modern medical research emphasizes the rapid translation of basic research into clinical applications, which is driven by the need to meet the demands of medical practice. Knowledge Advancement and Economic and Community Benefits, the two types of contributions, often do not occur simultaneously. Knowledge Advancement typically involves basic research and theoretical innovation, with a primary focus on the academic sphere. In contrast, economic and community benefits are often derived from applied research. There is a gap between basic research and the generation of significant economic and societal benefits. Additionally, there are inherent differences in research goals between basic and applied research. These disparities in goals cause scientists to concentrate more on a single area when conducting scientific research, which reduces the possibility of making both kinds of contributions in one paper.

In terms of more specific contributions, the majority (64.20%) of the nine scientific research contributions appear independently. There are 899 papers (28.03%) that demonstrate distinct scientific research contributions simultaneously. At most, six types of contributions appear simultaneously, but there is only one such paper. This shows that a study usually focuses on a single aspect to make outstanding contributions, and multiple contributions are less likely to occur at the same time. Table 9 shows the pairwise co-occurrence of scientific research contributions. The numbers in the table represent the count of papers in which contributions co-occur, indicating how many papers possess both contributions simultaneously. A darker shade in a cell signifies a greater intensity of co-occurrence of contributions. Among them, the most frequently co-occurring scientific research contributions are "Insights & Findings" and "Clinical Management and Treatment" (436), followed by "Insights & Findings" and "Data & Methods" (197), "Data & Methods" and "Clinical Management and Treatment" (143), as well as "Clinical Management and Treatment" and "Healthcare Services" (125). This implies that papers are more likely to generate other kinds of contributions when they contribute to the fields of idea creation, data methodologies, or clinical management and therapy. "Public Health Policy" contribution occurs infrequently with other kinds of contributions; in other words, public health policy is a relatively independent contribution.

Table 9. Co-occurrence of types of scientific research contributions.

Contribution type	1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3
1.1	0	85	37	17	57	7	8	6	3
1.2	85	0	197	106	436	28	62	56	20
1.3	37	197	0	43	143	16	27	15	15
2.1	17	106	43	0	55	9	27	15	5
2.2	57	436	143	55	0	38	125	65	20
2.3	7	28	16	9	38	0	8	10	6
3.1	8	62	27	27	125	8	0	13	5
3.2	6	56	15	15	65	10	13	0	5
3.3	3	20	15	5	20	6	5	5	0

High RStar - Low Citation papers and Low RStar - High Citation papers

The analysis results show a significant positive correlation between RStar and Citation of papers recommended by H1 Connect. This section explores some exceptions underlying this correlation. We define papers with the RStar in the top 10% but Citation in the bottom 10% as "High RStar – Low Citation papers (HR - LC)," totaling 39. Papers with RStar in the bottom 10% but Citation in the top 10% are termed "Low RStar – High Citation papers (LR - HC)," amounting to 293. An analysis focused on these two groups of papers, covering RStar, Citation, AAS, publication years, paper types, and other relevant attributes, is conducted to preliminarily identify characteristics of papers where the level of reviewer recognition significantly differs from Citation. The results are presented in Figure 7. As shown in Figure 7(a), the topic distribution of the two specific sub-datasets is similar to that of the overall dataset, with *Neoplasms* having the largest proportion of papers and *Respiratory Tract Diseases* having the least. From the perspective of document types, as shown in Figure 7(b), Article is the main type, and it accounts for a larger proportion of the HR - LC papers. In terms of publication year, as shown in Figure 7(c), there is a big difference between the two specific sub-datasets, with HR - LC papers being published later, mostly in 2019 and 2020, while LR - HC papers are published earlier, since paper citations take time to accumulate. With a notable separation between the two, Figure 7(d) shows the distribution of AAS and Citation for HR - LC papers and LR -HC papers. HR -LC papers have an average AAS of just 5.25, with AAS values ranging from 1 to 48.1. LR -HC papers, on the

other hand, have an average AAS of 629.15, and their AAS values range between 5.25 and 10528.266. This means that compared to papers with greater RStar, those with higher citations typically garner more social attention. The social attention received by LR - HC papers is significantly higher than that received by HR - LC papers.

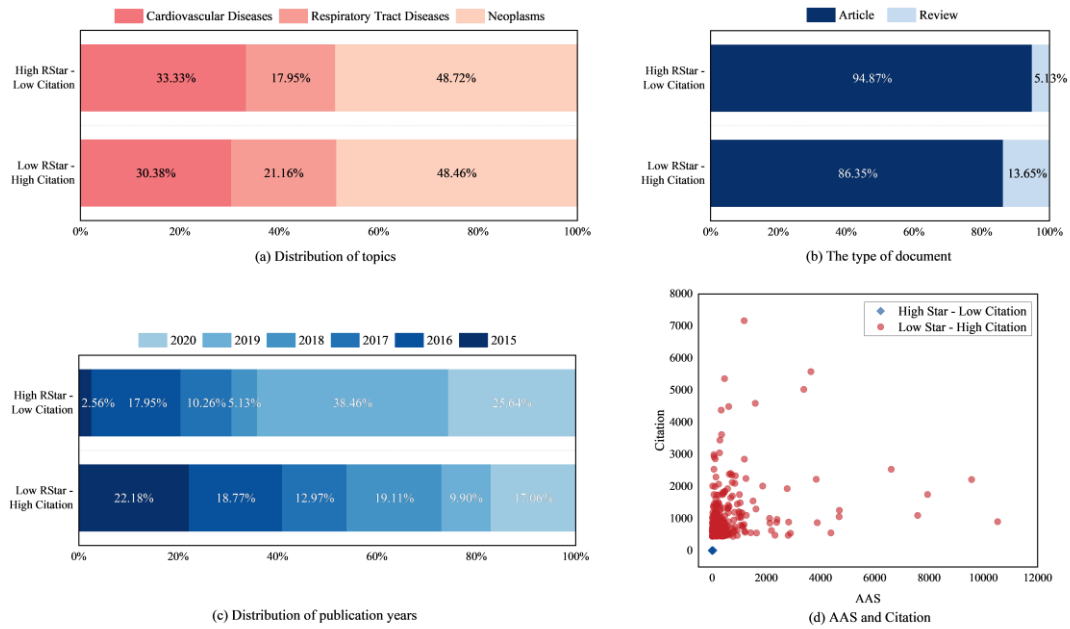


Figure 7. Characteristics of specific sub-datasets.

In the HR - LC and LR - HC papers, 21 papers (53.85%) and 57 papers (19.45%) respectively contained explicit scientific research contribution statements in their open peer review texts. The specific distribution is shown in Table 10. Similar to the overall distribution of scientific research contributions, "Insights & Findings" and "Clinical Management and Treatment" are the most common types of contributions. In the LR - HC papers, only these two types account for more than 10%, making them the dominant contributions. For the HR - LC papers, in addition to these two types, contributions related to "Data & Methods" are also notable. The scientific research contributions of HR - LC papers are concentrated in specific areas, with no contributions related to "Clinical Trial Outcomes," "Healthcare Services," or "Public Health Policy." In contrast, LR - HC papers address contributions across all nine types. This suggests that papers with greater contributions to economic and

community benefits tend to receive higher Citation and lower RStar, while papers focused more on theoretical innovation and clinical applications, with fewer contributions to economic and community benefits, often receive higher RStar but lower Citation.

Table 10. Distribution of scientific research contributions of special sub-datasets.

	1.1	*1.2*	*1.3*	*2.1*	*2.2*	*2.3*	*3.1*	*3.2*	*3.3*
High RStar - Low Citation	5.56 %	25.00 %	22.22 %	8.33 %	36.11 %	0.00 %	0.00 %	2.78 %	0.00 %
Low RStar - High Citation	6.25 %	46.25 %	2.50%	5.00 %	26.25 %	2.50 %	5.00 %	1.25 %	5.00 %

Discussion & conclusion

This study analyzed peer review data and impact indicators of papers on *Cardiovascular Diseases*, *Respiratory Tract Diseases*, and *Neoplasms*, revealing significant differences in the distribution of relevant indicators among different topics. At the same time, a significant correlation between peer review data and impact indicators was verified. Additionally, the study found that the scientific research contribution types of the papers exhibited clustering. The validity and reliability of open peer review data have been somewhat confirmed by this study, which also offers helpful references for better application of peer review data in academic evaluation practice. The results of peer review judge the value of a paper from the perspective of experts, while traditional citation and altmetrics consider the quality and influence of a paper from the perspective of scholars and the public. These indicators all play an important role in scientific evaluation. These three evaluation methods complement each other and together provide a strong basis for the evaluation of scientific research outcomes.

In terms of topic differences, this study conducted a statistical analysis of papers on *Cardiovascular Diseases*, *Respiratory Tract Diseases*, and *Neoplasms*. The analysis revealed differences in RNumber, RStar, Citation, and AAS among the papers in these three topics, indicating that the performance of papers across different evaluation perspectives is influenced by the research topic. Further pairwise

comparisons of the topics revealed that there were statistically significant differences between some topics ($P<0.05$), which highlights the need to consider the characteristics and priorities of different research fields when establishing the scientific research evaluation system. For example, *Cardiovascular Diseases* may focus more on clinical outcomes and the impact on healthcare services, while *Neoplasms* may be evaluated based on its contribution to drug development as well as clinical management and treatment. By adopting differential evaluation criteria for specific topics, with each topic being assessed based on its unique aspects, the academic evaluation system can more accurately capture the true contribution and value of research in different fields.

In terms of the relationship between indicators, this study conducted Spearman correlation analysis to explore the relationship between open peer review indicators and impact indicators. The results showed significant positive correlations among RNumber, RStar, Citation, and AAS. Peer review indicators, along with Citation and AAS evaluate scientific research from different perspectives, with varying emphases. This suggests that peer review data and impact indicators should complement each other in research evaluation. The consistency observed also indicates that peer review is an effective scientific evaluation method. Furthermore, compared to the delayed nature of citation, open peer review can help predict a paper's impact and identify valuable research with greater potential for academic and social impact after publication.

In terms of scientific research contributions, the study found that most papers recommended by H1 Connect tend to focus more prominently on one specific area of contribution, and the probability of multiple contributions occurring is relatively low.. The findings also reveal that while there are slight differences in the distribution of contribution types among the three topics, most research papers primarily focus on advancing insights and findings or contributing to clinical management and treatment. This indicates that in the field of biomedicine, academic research plays a crucial role not only in advancing the boundaries of disciplines and expanding knowledge systems, but also in optimizing clinical decision-making, improving treatment strategies, and ultimately enhancing public health outcomes. Another important finding is that when papers have contributions in viewpoint discovery, data methods, or clinical management and treatment, they are more likely to trigger other types of contributions, while contributions related to public health policy less frequently co-occur with other types of contributions. This suggests that there remains a gap between biomedical research findings and the translation into policy. Papers in the biomedical field often focus on theoretical innovation, technological

breakthroughs, or clinical applications, typically centered on specific diseases. The development of public health policies requires not only scientific evidence but also a comprehensive consideration of factors such as implementation challenges, economic costs, and other multifaceted aspects. Consequently, this highlights the need for a more comprehensive approach to evaluating scientific research, one that accounts for the diversity of contributions across different research areas. Rather than relying solely on a single indicator, academic evaluation should incorporate multiple indicators to reflect a paper's contributions across various dimensions.

In summary, the results of this study highlight the value of peer review in academic evaluation. In practice, it is crucial to recognize the multiple contributions research can make and consider the unique characteristics of different research fields. A more comprehensive and diversified academic evaluation system, which should include impact indicators and peer review data, will better capture the multifaceted nature of scientific contributions. As research fields continue to evolve and become increasingly specialized, the evaluation system must adapt to ensure that it accurately reflects the diversity and influence of scientific work. Thus, it can promote a more open and comprehensive academic evaluation process.

There are also some limitations in this study. Due to the characteristics of the H1 Connect platform, the data samples selected in this study belong to the field of biomedicine, and there may be differences between different topics. In the future, the scope of the research can be further expanded to other fields to validate the generalizability of the conclusions. In addition, the peer review process is affected by multiple factors, such as the reviewer's research interests. In the future, other dimensions can be supplemented to explore the differences in peer review behavior under the influence of multiple factors.

References

Bravo, G., Farjam, M., Grimaldo Moreno, F., Birukou, A., & Squazzoni, F. (2018). Hidden connections: Network effects on editorial decisions in four computer science journals. *Journal of Informetrics, 12*(1), 101–112.

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., & Squazzoni, F. (2019). The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications, 10*(1), 322.

Chen, C., Cheng, Z., Wang, C., & Li, L. (2023). Identification and utilization of key points of scientific papers based on peer review texts. *Journal of the China Society for Scientific and Technical Information, 42*(5), 562–574.

Cheng, X., Wang, H., Tang, L., Jiang, W., Zhou, M., & Wang, G. (2024). Open peer review correlates with altmetrics but not with citations: Evidence from Nature Communications and PLoS One. *Journal of Informetrics*, *18*(3), 101540.

Demarest, B., Freeman, G., & Sugimoto, C. R. (2014). The reviewer in the mirror: Examining gendered and ethnicized notions of reciprocity in peer review. *Scientometrics*, *101*(1), 717–735.

Fox, C. W., & Paine, C. E. T. (2019). Gender differences in peer review outcomes and manuscript impact at six journals of ecology and evolution. *Ecology and Evolution*, *9*(6), 3599–3619.

Fromm, M., Faerman, E., Berrendorf, M., Bhargava, S., Qi, R., Zhang, Y., Dennert, L., Selle, S., Mao, Y., & Seidl, T. (2021). Argument Mining Driven Analysis of Peer-Reviews. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(6), Article 6.

Ghosal, T., Kumar, S., Bharti, P. K., & Ekbal, A. (2022). Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, *17*(1), e0259238.

Ghosal, T., Verma, R., Ekbal, A., & Bhattacharyya, P. (2020). DeepSentipeer: Harnessing sentiment in review texts to recommend peer review decisions. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*,1120–1130.

Han, R., Zhou, H., Zhong, J., & Zhang, C. (2022). Characterizing Peer Review Comments of Academic Articles in Multiple Rounds. *Proceedings of the Association for Information Science and Technology*, *59*(1), 89–99.

Hua, X., Nikolov, M., Badugu, N., & Wang, L. (2019). *Argument mining for understanding peer reviews*. *1*, 2131–2137.

Lin, Y., Wang K., Ding K., & Xu, K. (2021). Quantitative research on qualitative evaluation of academic papers. *Information Studies:Theory & Application*, *44*(8), 28–34.

Luo, Z., Cai, L., Qian, J., & Lu, W. (2021). Research on the recognition of innovative contribution sentences of academic papers. *Library and Information Service*, *65*(12), 93–100.

Ni, J., Zhao, Z., Shao, Y., Liu, S., Li, W., Zhuang, Y., Qu, J., Cao, Y., Lian, N., & Li, J. (2021). The influence of opening up peer review on the citations of journal articles. *Scientometrics*, *126*(12), 9393–9404.

Qin, C. (2020). Exploring the distribution regularities of referees' comments in IMRAD structure of academic articles. *18th International Conference on Scientometrics and Informetrics, ISSI 2021*, 1527–1528.

Shi, Z., Zhu L., & Le, X. (2024). Material Information Extraction Based on Local Large Language Model and Prompt Engineering. *Data Analysis and Knowledge Discovery*, *8*(7), 23–31.

Wang, K., & Wan, X. (2018). Sentiment Analysis of Peer Review Texts for Scholarly Papers. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 175–184.

Wang, P., Hoyt, J., Pöschl, U., Wolfram, D., Ingwersen, P., Smith, R., & Bates, M. (2016). The last frontier in open science: Will open peer review transform scientific and scholarly publishing? *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–4.

Wang, P., Williams, J., Zhang, N., & Wu, Q. (2020). F1000Prime recommended articles and their citations: An exploratory study of four journals. *Scientometrics*, *122*(2), 933–955.

Wang, H. (2023). Value, mechanism and strategies of open peer review for periodicals. *Acta Editologica*, *35*(2), 147–151.

Wolfram, D., Wang, P., Hembree, A., & Park, H. (2020). *Scientometrics*, *125*(2), 1033–1051.

Xie, W., Jia, P., Zhang, G., & Wang, X. (2024). Are reviewer scores consistent with citations? *Scientometrics*, *129*(8), 4721–4740.

Zhang, G., Wang, L., Xie, W., Shang, F., Xia, X., Jiang, C., & Wang, X. (2021). "This article is interesting, however": Exploring the language use in the peer review comment of articles published in the BMJ. *Aslib Journal of Information Management*, *74*(3), 399–416.

Zong, Q., Fan, L., Xie, Y., & Huang, J. (2020). The relationship of polarity of post-publication peer review to citation count: Evidence from Publons. *Online Information Review*, *44*(3), 583–602.

Zong, Q., Xie, Y., & Liang, J. (2020). Does open peer review improve citation count? Evidence from a propensity score matching analysis of PeerJ. *Scientometrics*, *125*(1), 607–623.

https://doi.org/10.51408/issi2025_048

From Search to Recommendation: Using an LLM to Assess the Usefulness of Academic Articles

Frans van der Sluis[1], Alesia Zuccala[2], Haakon Lund[3]

[1] *frans@hum.ku.dk*, [2]*a.zuccala@hum.ku.dk*, [3]*hl@hum.ku.dk*
Department of Communication, University of Copenhagen, Karen Blixens Vej 8, 2300, Copenhagen (Denmark)

Abstract

The goal of this paper is to explore what an AI-powered LLM can do to help academics/scientists organize, classify, summarize, and make recommendations concerning the relevance of reference articles for the preparation of a literature review. Literature reviewing is a core task in academia, which requires systematic planning and thinking, but today, enormous amounts of information make this process onerous. Many scholars are familiar with research management tools like Endnote, Zotero and Mendeley; however, the advent of LLMs means that new potentialities are on the horizon. We investigate one LLM's ability to make synthesized judgements about a set of article abstracts retrieved from Scopus (n=194), to prepare a literature review for one 'case paper'. Our finding was that its selecting and filtering capabilities were not quantitatively impressive, though qualitatively, it produced many useful recommendations. Here, we describe the kind of inferences the LLM can make about scientific relevance and discuss the potential of LLMs in utilizing academic literature.

Introduction

As the amount of scientific literature published each year increases, it becomes harder to keep up to date with current information and use it when writing a manuscript. Despite this challenge, literature search and reviewing are core skills that an academic needs to situate and contextualize new work. For the uninitiated, Onwuegbuzie and Frels (2015) have produced a guidebook, titled: S*even Steps to A Comprehensive Literature Review*. Within this 'seven step' approach, the first five constitute an "exploratory phase," followed by an 'interpretive phase,' then finally the seventh 'communicative phase.' The exploratory phase alone involves: 1) establishing a research question, 2) initiating a systematic search, 3) storing, and organizing the information, selecting, and then 'deselecting' information based on an established set of criteria – i.e., to choose or not to choose a source.

Early on, this may have involved basic note cards, but today, the average graduate student does not have to sit amongst notes and papers "Piled high" and "Deep" (PhD) just to produce a comprehensive literature review. Today, software tools like Endnote, Mendeley, and Zotero, make this process much easier. All three tools are useful for storing and organizing references, keeping user notes, inserting citations into a manuscript, and automatically formatting bibliographies. An added benefit of Mendeley and Zotero, is that both possess capabilities as reference finders. For example, the Mendeley 'suggest' feature, implements several different recommendation algorithms (i.e., collaborative & content-based filtering; popularity-based & trend-based models) to help academics "discover new research"

based on [their libraries, search behaviors], and general short-term and long-term interests" (Wordpress, 2015).

Many academics are familiar with Mendeley (Zaug et al., 2011), but now AI-powered Large Language Models (LLM) are inspiring researchers to investigate how useful they are at producing textual summaries (Ahmed Antu et al., 2023, Cai et al., 2024; Jin et al., 2024; Nechakhin,2024), as well as feedback (Liang et al., 2024). Experts are positive about the range of applications and potential impact that LLM will have on the higher education system (Pearson, 2024; Luo et al., 2025). Still, they advise academics to maintain skills in critical thinking, problem solving, and ethical decision-making (Fetcher et al., 2023; Watson et al., 2025).

In this paper we look to an LLM both as a tool for filtering relevant research articles and for recommending how useful the articles are for preparing a literature review. We address this specifically by retrieving document abstracts from Scopus and prompting an LLM to sort and contextualize them for their potential value as references, beyond mere keyword or topical relatedness. This is challenging, since current academic search engines (e.g., Scopus) already deliver relevant results based on extensive queries and keyword-based retrieval, making it difficult to improve significantly upon their effectiveness. In contrast to existing recommender systems, our case serves a well-defined need for academic literature in relation to a paper in progress. To ensure accessibility and scalability, we use an open-source LLM and consumer-grade GPU.

Can an LLM help with literature review? Our aim is to answer this question in the context of academic search (Christou et al., 2024), with the added goal of extending earlier work (Azzopardi & Van Der Sluis, 2024; Van der Sluis & Azzopardi, 2025). Specifically, we examine how an LLM can estimate and detail the relevance and usefulness of scientific article abstracts for writing a 'case paper,' which builds on and follows from that earlier research.

Related work

A key issue in academic search is the subjective and multifaceted nature of relevance (Christou et al., 2024; Jordan & Tsai, 2024). Search engines like Google Scholar rely heavily on ranking algorithms that prioritize citation counts and the presence of search terms (Beel & Gipp, 2009; Mallapaty, 2024), but these methods are neither transparent nor comprehensive. This reliance on citation counts reinforces biases such as the "Matthew effect," where already-cited works gain disproportionate visibility, while less-cited but potentially valuable contributions are relegated to the "long tail" of academic literature (Gould, 2009). It also means that search engines may be misconstrued as informants in knowledge production, rather than inert sources of information. This results in a system that favors established viewpoints and overlooks innovative or niche research, limiting the diversity of knowledge accessible to researchers.

Current approaches to relevance evaluation, including binary and graded judgments, focus primarily on topicality or algorithmic matching, often failing to address a user's specific goals/needs or context (Borlund, 2003; Saracevic, 2007). While graded judgments offer a more nuanced assessment, by assigning degrees of relevance, they

remain centered on query-content relationships and fall short of addressing the practical value of information in specific tasks (Cole et al., 2009; Van der Sluis et al., 2010). The reliance on these relevance-based judgments underpins traditional search engine algorithms like Google Scholar, which conceptualize relevance as relatedness rather than usefulness. Judging usefulness, however, needs more information than can typically be captured in a query or easily evaluated on a search engine index (Cole et al., 2009).

Recent advancements in language models have led to the use of LLMs for judging the relevance and ranking of research papers (Luo et al., 2025). These developments are part of a broader suite of Retrieval-Augmented Generation (RAG) technologies, where LLMs interact with traditional search engines and indexes to ground their outputs in external, up-to-date knowledge sources (Argawal et al., 2022; Huang & Huang, 2024). RAG enables extensive, semantic queries that represent full abstracts when searching a database. Here, LLMs are used for query expansion by extending abstracts with related terms and pseudo-references, leveraging information available in the corpus (Shi et al., 2023). Additionally, LLMs assist in relevance estimation and re-ranking, using both supervised and zero-shot methods to reorder search results based on their conceptual fit with an abstract (Argawal et al., 2024; Hou et al., 2023). Their ability to query by abstract enhances literature exploration by understanding context beyond simple keyword matching, allowing for more precise, user-specific retrieval.

Despite these advancements, the primary focus of LLMs in academic research has remained on generation rather than retrieval, particularly in summarization and literature review writing (Pearson, 2024; Luo et al., 2025). Existing systems, such as AutoCite (Wang et al., 2021) and BACO (Ge et al., 2021), generate structured citation texts by leveraging citation networks and textual data to produce contextually relevant citation texts. Similarly, hierarchical clustering techniques in RAG-based models enhance literature reviews by structuring research fields. These systems excel at summarization and organization, enabling automated literature review writing. While advances in sentence-based planning and contextual summarization have refined the automated presentation of prior work, no existing system explicitly supports ideation and writing by helping authors strategically select and integrate references. In this work, we take a step before fully automated writing, exploring whether an LLM can assist authors in assessing a reference's contribution to their own work.

Method

Instruments and Equipment

The Gemma2 language model was used[1], an open-source large language model (LLM) developed by Google, which features 27 billion parameters. The model was

[1] URL: https://huggingface.co/bartowski/gemma-2-27b-it-GGUF
Model file: gemma-2-27b-it-Q6_K_L.gguf

instruction-trained and employed a recommended quantization level of 6 bits (Team et al., 2024). Gemma2 represents a trend towards smaller yet high-performing models, designed for open exploration, fine-tuning, and testing in diverse applications[2]. Instruction-tuned models usually follow a system-user-assistant prompt structure. Gemma2 omits the system role but supports an assistant role for examples; however, this was deliberately omitted to focus on user-directed instructions.

Inferencing was performed using a consumer-grade Nvidia RTX 4090 GPU, equipped with 24 GB of VRAM. The model's context window was set to 2024 tokens to fit within the available VRAM. This limits the number of tokens that can be included in a single prompt, restricting the number of abstracts that can be supplied simultaneously. LlamaCPP, a foundational API for LLMs, was used to structure prompts and ensure compatibility with the Gemma2 model.

Procedure and Materials

The procedure had two phases (see Figure 1). First, we iteratively refined a Scopus query to identify search results relevant to the case paper's topics: information seeking and green consumption (Azzopardi & Van Der Sluis, 2024; Van der Sluis & Azzopardi, 2025). Scopus, a comprehensive database of academic literature (Mallapaty, 2024), provides detailed results, including titles, abstracts, authors, and other metadata. Standard keyword selection and refinement practices focused on the query while limiting the results list's size. An abstract of the case paper informed query development, with synonyms generated using ChatGPT 4o and selectively added to avoid overexpanding the results list. This process ensured a highly relevant set of abstracts. The results are available on Github (https://github.com/fsluis/scopus-llm-review). In total, 194 abstracts were obtained. The final Scopus query was:

TITLE-ABS-KEY(("search behavior" OR "search behaviour" OR "information seeking" OR "web search" OR "information evaluation" OR "information retrieval" OR "consumer search behavior" OR "green complexities" OR "search on information" OR "greenwashing" OR "green washing" OR "information barriers" OR "knowledge barriers") AND ("responsible consumption" OR "sustainable consumption" OR "green consumerism" OR "conscious consumer" OR "ecological consumer" OR "environmentally sustainable" OR "eco-conscious" OR "ethical consumerism" OR "ethical consumer" OR "socially responsible purchasing" OR "sustainable behavior" OR "sustainable behaviours" OR "sustainable decision making" OR "eco-friendly decision-making" OR "consumer decision making" OR "ethical decision making" OR "purchase decision making" OR "green consumption" OR "green shopping" OR "consumption gap") AND NOT ("infrastructure" OR "enterprise" OR "corporate"))

Final Scopus query

[2] For an informal benchmark, visit https://dubesor.de/benchtable

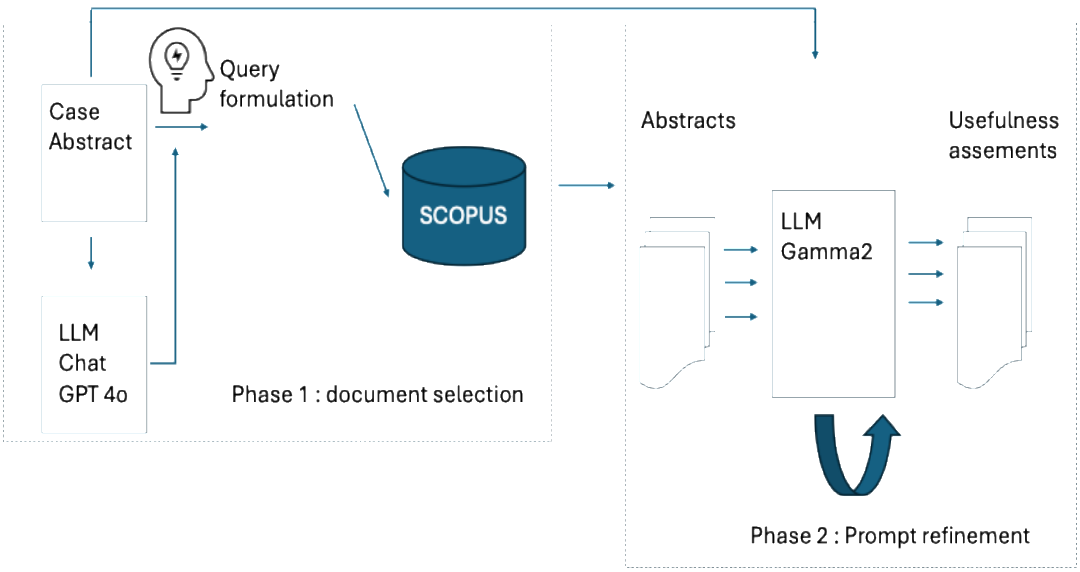


Figure 1. Document selection and prompt refinement. Phase 1 includes an LLM to support manual refinement of a Scopus query. Phase 2 includes an LLM for automated usefulness assessments of article abstracts retrieved from Scopus.

Table 1. Prompt used with Gemma 2.

Section	Prompt
1	I want you to evaluate whether an abstract of a reference paper is relevant to a paper I'm writing. I'll give you details of both my paper and the reference paper.
2	My paper: Abstract: {my_abstract} Reference paper: Title: {title} Abstract: {abstract}
3	I am particularly interested in knowing whether a paper relates to either of: a) Information seeking: Studies of information seeking and sustainable or responsible consumption, including information seeking challenges experienced by consumers; b) Information availability: Studies showing the influence of information availability or barriers on responsible or sustainable consumer behavior; c) Asymmetries: Studies showing the existence of information asymmetries between market players and consumers, such as through greenwashing practices; d) Sustainability: Studies showing the importance of sustainable practices, but are not directly relevant to my study; e) Other: There might be other categories of relations. Do feel free to add / interpret new types of relations.

4	Be critical when estimating relevance. If it is not about sustainability or responsible consumption, it is not relevant.
5	Does the reference paper seem relevant? If yes, how can it be utilized in my research? Answer in a structured way: Relevance: Yes, possibly, no Relation: Seeking, availability, asymmetries, sustainability, other Utilization: Explain how this paper can be utilized in my research

Results

Here we analyze the responses received from the LLM, both quantitatively and qualitatively. The focus is purely on the end-product of our exploration. The complete set of responses is available on GitHub (https://github.com/fsluis/scopus-llm-review).

Quantitative results

The LLM did not reduce the amount of text needed for an evaluation from a quantitative perspective. The average text length of the full-text assessments is 1463.51 characters, while abstracts, at 1490.07 characters on average, are close to the same length. Of the 194 references analyzed, 41 were identified as relevant, while 129 were labeled 'possibly' relevant. With only 24 reference articles identified as not relevant, the tool did not drastically reduce what was considered useful as a final collection. In fact, it left a substantial amount of work for the authors to manually sort through and identify relevant versus non-relevant articles. The usefulness of the LLM tool was, from a quantitative perspective, limited.

Qualitative results

Qualitatively, we observed certain features of LLM responses that made it overall a powerful tool. Here, we highlight features of responses identified through initial observations and subsequent collaborative discussions amongst the authors. Quotes were selected randomly, though with a preference for brevity. The LLM responses were consistent in focus and style, suggesting that a sample of responses provides a reliable representation of the overall output.
Some of the full-text responses focused primarily on *relevance relations* specified in Section 3 of the prompt, expanding on the structured (quantitative) responses:

(1) *"**Information Availability:** The reference paper directly addresses how retailers can influence consumer behavior by making eco-friendly products more visible and accessible. This aligns with your focus on "information barriers" and how accessible information can drive prioritization of responsible aspects. You can cite this paper to support your argument that improving information availability is crucial for fostering responsible consumption."*

[DOI: 10.1016/j.jretconser.2016.07.008, Relevance: yes, Relation: availability, asymmetries]

(2) *"**Greenwashing as an Information Asymmetry:** The reference paper highlights how greenwashing practices create an information asymmetry between companies and consumers. Companies may present their products as more environmentally friendly than they actually are, misleading consumers who lack complete information."*

[DOI: 10.1108/YC-10-2020-1241, Relevance: yes, Relation: asymmetries]

These examples demonstrate that the LLM typically explains the relevance relation observed. Additional information is included from the reference paper abstract, as seen in Quote 2, and in certain cases about potential use cases, as seen in Quote 1. More often, the usage-part of the response expands on the initial relevance relations, by highlighting other potential uses:

(3) *"**Connecting to Your Research:** You can use this paper to support your argument about the importance of **information availability and transparency** in shaping responsible consumption. The reference paper demonstrates how a lack of accurate information (due to greenwashing) can hinder consumers' ability to make ethical choices, even if they have initial intentions to consume responsibly."*

[DOI: 10.1108/YC-10-2020-1241, Relevance: yes, Relation: asymmetries]

(4) *"**Establishing a baseline:** The paper's findings on the increasing prevalence of online information seeking for major financial decisions could be used to highlight the growing importance of online information for consumers in general. You could then argue that this trend extends to responsible consumption decisions, even if the paper itself doesn't directly address sustainability."*

[DOI: 10.1080/16522354.2007.11073450, Relevance: possibly, Relation: information seeking]

Quote 3 corresponds to previously presented Quote 2 [i.e., DOI: 10.1108/YC-10-2020-1241]. Here, the LLM's response directly supports a core argument of the case paper by identifying how greenwashing hinders sustainable consumption. Quote 4 indirectly supports a broader, non-core argument, concerning the importance of information seeking in consumer decision-making. Both illustrate the LLM's ability to expand on initial relevance assessments and describe how each paper might potentially be used, as requested in Section 5 of the prompt.

The LLM also identified several 'other' relationships, peripheral to the case paper. 'Other' relationships were less common (only 16 occurrences with a positive relevance annotation) and typically came with more extensive full-text responses:

(5) *"**Information Seeking in Niche Communities:** The paper highlights how online drug forums serve as platforms for information exchange about a specific product (MXP). This could be drawn upon to illustrate how consumers in niche markets, even those involving potentially harmful products, actively seek information to inform their purchasing decisions. You could discuss this*

as a parallel to consumers seeking information about ethically sourced or sustainable products."

[DOI: 10.1080/02791072.2014.974002, Relevance: possibly, Relation: other]

(6) *"**Decision-Making Strategies**: The paper identifies decision-making strategies used by consumers in omni-channel contexts. Understanding these strategies could be helpful in contextualizing how information about responsible consumption influences consumer choices. For example, if consumers primarily rely on "weighted additive rule" (WADD), it suggests they weigh different product attributes, potentially including ethical considerations, in a systematic manner."*

[EIC: 2-s2.0-85126569618, Relevance: possibly, Relation: other]

Quote 5 focuses on the consumption of a synthetic drug (MXP) with potential personal harm rather than environmental harm. Quote 6 focuses on fashion buying, suggesting that similar heuristics may apply to sustainable considerations. Both examples provide interesting ways to enrich our case paper, methodologically and/or contextually.

Sometimes the LLM provided a few creative (but good!) suggestions for alternative papers to use; papers that did not have any initial relevance relation to the case paper, but may be useful when writing the implications or discussion sections of the case paper:

(7) *"**Expanding Your Discussion:** You could discuss the reference paper findings in the context of your own research on information barriers. provides a concrete example of how information asymmetries can act as barrier to responsible consumption. You could also explore how you proposed solutions (improved information systems, decision-support tools could help mitigate the negative effects of greenwashing and empowe consumers to make more informed decisions."*

[DOI: 10.1108/YC-10-2020-1241, Relevance: yes, Relation: asymmetries

(8) *"**Exploring interventions:** The study's use of retro-cueing to increas information availability could inspire you to explore potential intervention or design strategies that make ethical information more salient and accessibl to consumers."*

[DOI: 10.1016/j.jbusres.2019.08.012, Relevance: possibly, Relation: availability

Both of these examples point at possible use cases beyond the initial relevance relations prompted. It expands the scope of literature considered, surfacing sources that might otherwise be overlooked.

And finally, in addition to reference papers with potentially creative uses, the LLM responded with critical notes, related to Section 4 of the prompt:

(9) *"Remember to critically evaluate the reference paper's methodology and findings in the context of your own research."*

[DOI: 10.1080/17543266.2024.2343934, Relevance: yes, Relation: Asymmetries, Sustainability]

*"**Important Note:** While the reference paper offers relevant insights, it's crucial to acknowledge the differences in context. Health information seeking and responsible consumption involve distinct decision-making processes and motivations. Therefore, avoid directly applying the findings to your research without careful consideration and contextualization."*

[DOI: 10.1016/j.hlpt.2024.100842, Relevance: possibly, Relation: information seeking]

These critical notes appeared frequently. Quote 9 reminds us that abstracts alone are insufficient to assess a reference paper's merits. Quote 10, which was more common, cautions against over-generalizing a reference paper's findings to the case paper's context. We found these critical notes to be well-grounded and comment on this further in the Discussion section.

Overall, these 10 quotes highlight the LLM's strength in contextualizing reference abstracts and presenting structured, clear assessments. Clear headings enabled quick scanning of reference papers, while the structured and concise format made it easier to evaluate papers on their potential usefulness. By going beyond relevance to provide actionable suggestions, some responses guided the incorporation of references into the case paper, helping refine arguments and expand its scope and implications.

Discussion

This work positions LLMs as a transformative tool in literature reviews by addressing two key contributions. First, it demonstrates how LLMs fulfill the longstanding ambition of implementing usefulness as a core relevance concept, moving beyond traditional binary or graded relevance judgments to actionable insights. By structuring responses with relevance labels and task-specific suggestions, LLMs bridge the gap between search engine outputs and the practical support of ideation and writing processes. Second, it extends the scope of Retrieval-Augmented Generation (RAG) approaches, showing that information retrieval not only enhances text generation, but also that LLMs can augment traditional article-based approaches. By connecting relevance to usefulness, LLMs unify these two paradigms, advancing both the practical application of retrieved items and raising the possibility of generation-augmented retrieval (GAR), where LLMs become part of the retrieval process.

Our results show that LLM-generated assessments add significant value beyond reference abstracts by helping researchers interpret diverse and dispersed details. By consolidating information into structured insights, LLMs assist in evaluating both relevance and usefulness in relation to a researcher's work. This streamlines the literature review process in two key ways: saving time when sifting through large volumes of references and supporting writing through creative ideas and recommendations for integrating citations. For researchers with limited time or resources, LLMs running on consumer-grade hardware provide a scalable and efficient alternative to traditional methods. However, these findings are based on a

sole case study and reflect the authors' perspectives, which may limit their generalizability. Even though both the authors and intended readership are well-positioned to judge the examples presented in this case study, it remains an open question as to whether these conclusions hold across different authors, disciplines, or research contexts.

These findings suggest a broader role for LLMs in the literature review cycle. By mitigating biases introduced by citedness-based rankings in search engines like Google Scholar (Mallapaty, 2024), LLMs can delve into less-explored references, potentially democratizing the academic literature (Fecher et al., 2023). By easing access to lesser-cited but valuable works, LLMs could even out the long tail of underused articles and give smaller, lesser-known studies a better chance of being cited. This contributes to a more equitable distribution of academic attention and resources.

Despite these advantages, quality control remains a critical limitation. The risk of misuse, where LLMs might shortcut the review process without proper validation, underscores the need for robust quality mechanisms. In a landscape where LLMs increasingly support both the reading and writing of academic literature (Fecher et al., 2023), the emphasis on peer review and expert judgment is heightened. This is especially vital given the proliferation of non-peer-reviewed repositories like arXiv and the potential for errors to propagate, echoing concerns seen in the replication crisis within other disciplines (Open Science Collaboration, 2015). As reliance on LLMs grows, the importance of quality control (Van der Sluis, 2022; Van der Sluis et al., 2024) cannot be overstated.

Future work

To generalize the current findings, future work could consider repeating the presented approach across different case papers and disciplines. Establishing ground truth labels would allow for more formal evaluation using retrieval metrics such as precision and recall, while also enabling comparative testing against existing tools such as Scopus rankings or Zotero Suggest. This could help quantify the practical advantages of LLMs in literature review workflows, beyond the currently highlighted qualitative strengths such as interpretability and perceived usefulness.

Future work could also extend the technical contributions of this study by testing different LLMs and refining prompt design. In addition, automated querying, developed and researched as part of the RAG suite, presents promising opportunities for academic literature search. LLMs can support query drafting, refinement, and synonym generation for complex academic search engines like Scopus. While the current study focuses on interpreting retrieved abstracts, future systems could integrate both querying and evaluation in a single LLM workflow.

Nevertheless, the computational demands and environmental footprint of LLMs warrant continued investigation. Developing efficient workflows for consumer-grade hardware could broaden access and promote more sustainable and responsible deployment of these tools in academic research. Addressing these challenges alongside optimizing consumer-grade hardware use offers a dual opportunity: advancing LLM capabilities for academic purposes and promoting their responsible,

sustainable deployment. These efforts would support more equitable access to research tools, reinforcing the democratization of academic practices.

References

Ahmed Antu, S., Chen, H., & Richards, C. K. (2023). Using LLM (Large Language Model) to improve efficiency in literature review for undergraduate research. In *Proceedings of the Workshop on Empowering Education with LLMs – the Next-Gen Interface and Content Generation*, Tokyo, Japan, 7 July 2023, 8–16.

Agarwal, S., Laradji, I. H., Charlin, L., & Pal, C. (2024). LitLLM: A toolkit for scientific literature review. *arXiv*, doi:10.48550/arXiv.2402.01788.

Aytar, A. Y., Kilic, K., & Kaya, K. (2024). A retrieval-augmented generation framework for academic literature navigation in data science. *arXiv*, doi:10.48550/arXiv.2412.15404.

Azzopardi, L., & Van der Sluis, F. (2024). Seeking socially responsible consumers: Exploring the intention–search–behaviour gap. In P. Clough, M. Harvey, & F. Hopfgartner (Eds.), *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24)* (pp. 153–164), Sheffield, UK, March 10–14, 2024. ACM. doi:10.1145/3627508.3638324

Van der Sluis, F., & Azzopardi, L. (2025). Search changes consumers' minds: How recognizing gaps in understanding drives ethical choices. In D. McKay & D. Oard (Eds.), *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25)*, Melbourne, Australia, March 24–28, 2025. ACM. doi:10.1145/3698204.3716456

Beel, J., & Gipp, B. (2009). Google Scholar's ranking algorithm: The impact of citation counts (an empirical study). In *Proceedings of the 2009 Third International Conference on Research Challenges in Information Science (RCIS)* (pp. 439–446). Fez, Morocco, April 22–24, 2009. IEEE. doi:10.1109/RCIS.2009.5089308

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10), 913–925. doi:10.1002/asi.10286

Cai, H.; Cai, X.; Chang, J.; Li, S.; Yao, L.;Wang, C.; Gao, Z.; Li, Y.; Lin, M.; Yang, S.; et al. (2024)

SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis. arXiv, doi: 10.48550/arXiv.2403.01976.

Christou, E., Parmaxi, A., & Zaphiris, P. (2025). A systematic exploration of scoping and mapping literature reviews. *Universal Access in the Information Society*, 24(1), 941–951. doi:10.1007/s10209-024-01120-3

Cole, M., Liu, J., Bierig, R., Gwizdka, J., Liu, C., Zhang, J., & Zhang, X. (2009). Usefulness as the criterion for evaluation of interactive information retrieval. In *Proceedings of the 2009 HCIR Workshop on Human–Computer Interaction and Information Retrieval* (pp. 1–4), Cambridge, MA, USA, October 23, 2009.

Fecher, B., Hebing, M., Laufer, M., Pohle, J., & Sofsky, F. (2025). Friend or foe? Exploring the implications of large language models on the science system. *AI & Society*, 40(2), 447–459. doi:10.1007/s00146-023-01791-1

Ge, Y., Dinh, L., Liu, X., Su, J., Lu, Z., Wang, A., & Diesner, J. (2021). BACO: A background knowledge- and content-based framework for citing sentence generation. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL/IJCNLP 2021)* (pp. 1466–1478). Online, August 1–6, 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.116

Gould, T. H. P. (2009). The future of academic publishing: Application of the long-tail theory. *Publishing Research Quarterly*, 25(4), 232–245. doi:10.1007/s12109-009-9134-y

Huang, Y., & Huang, J. (2024). A Survey on Retrieval-Augmented Text Generation for Large Language Models. *ArXiv*. doi: doi:10.48550/arXiv.2404.10981

Hou, Y.,Zhang, J.,Lin, Z., Lu, H., Xie, R., McAuley, J., Zhao, W.X. (2023) Large Language Models are zero-shot rankers for recommender systems. *ArXiv*, doi:10.48550/arXiv.2305.08845

Jin, H.; Zhang, Y.; Meng, D.; Wang, J.; Tan, J. (2024). A Comprehensive Survey on Process-Oriented Automatic Text Summarization with Exploration of LLM-Based Methods. *arXiv*, doi: doi:10.48550/arXiv.2403.02901.

Jordan, K., & Tsai, S. P. (2024). Ranking 'by relevance' in academic literature searches: Prevalence, definitions, and implications. *Postdigital Science and Education*. doi:10.1007/s42438-024-00530-z

Latif, E.; Fang, L.; Ma, P.; Zhai, X. (2023). Knowledge distillation of llm for education. *arXiv*, doi: 10.48550/arXiv.2312.15842.

Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D.; Yin, Y.; et al. (2023). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *arXiv*, doi:10.48550/arXiv.2310.01783.

Luo, Z., Yang, Z., Xu, Z., Yang, W., & Du, X. (2025). LLM4SR: A Survey on Large Language Models for Scientific Research. *ArXiv*. doi:10.48550/arXiv.2501.04306.

Mallapaty, S. (2024). Can Google Scholar survive the AI revolution? *Nature*, 635(8040), 797–798. doi:10.1038/d41586-024-03746-y

Nechakhin, V., D'Souza, J., & Eger, S. (2024). Evaluating large language models for structured science summarization in the open research knowledge graph. *Information(Switzerland), 15(6)*, 328. doi:10.3390/info15060328

Onwuegbuzie, A. J., & Frels, R. K. (2015). *Seven steps to a comprehensive literature review: A multimodal and cultural approach*. SAGE Publications.

Open Science Collaboration. (2015). *Estimating the reproducibility of psychological science*. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716

Pearson, H. (2024). Can AI review the scientific literature—and figure out what it all means? Nature, 635(8038), 276–278. doi:10.1038/d41586-024-03676-9

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58(13), 1915–1933. doi:10.1002/asi.20682.

Shi, Z., Gao, S., Zhang, Z., Chen, X., Chen, Z., Ren, P., & Ren, Z. (2023). Towards a unified framework for reference retrieval and related work generation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5785–5799). Singapore: Association for Computational Linguistics. doi:10.18653/v1/2023.findings-emnlp.385.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., et al. (2024). Gemma 2: Improving Open Language Models at a Practical Size. *arXiv*. doi: doi:10.48550/arXiv.2408.00118.

Van der Sluis, F., van den Broek, E. L., & van Dijk, B. (2010). Information Retrieval eXperience (IRX): Towards a human-centered personalized model of relevance. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2010)* (pp. 322–325). Toronto, Canada, August 31 – September 3, 2010. IEEE. doi:10.1109/WI-IAT.2010.222

Van der Sluis, F. (2022). *A conversationalist approach to information quality in information interaction and retrieval*. In *CHIIR'22 Workshop on Information Quality in Information Interaction and Retrieval*. doi:10.48550/arxiv.2210.07296

Van der Sluis, F., Faure, J., & Homnual, S. P. (2024). An empirical exploration of the subjectivity problem of information qualities. *Journal of the Association for Information Science and Technology*, 75(7), 829–843. doi:10.1002/asi.24884.

Watson, S., Brezovec, E., & Romic, J. (2025). The role of generative AI in academic and scientific authorship: An autopoietic perspective. *AI & Society, 1-11*. doi:10.1007/s00146-024-02174-w

Wang, Q., Xiong, Y., Zhang, Y., Zhang, J., & Zhu, Y. (2021). AutoCite: Multi-modal representation fusion for contextual citation generation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)* (pp. 788–796). Virtual Event, Israel, March 8–12, 2021. ACM. doi:10.1145/3437963.3441739

Wordpress (2015). A practical guide to building recommender systems. From Algorithms to Product. Mendeley Suggest. Available at: https://buildingrecommenders.wordpress.com/2015/11/13/mendeley-suggest/.

Zaugg, H., West, R. E., Tateishi, I., & Randall, D. L. (2011). Mendeley: Creating communities of scholarly inquiry through research collaboration. *TechTrends*, 55(1), 32–36. doi:10.1007/s11528-011-0467-y

https://doi.org/10.51408/issi2025_014

Gender Differences in Research Methods: Insights from Chinese Humanities and Social Sciences PhD Dissertations

Xinyu Deng[1], Hui Xu[2], Zihui Li[3], Huiwen Bai[4], Lanfeng Ni[5], Chengzhi Zhang[6]

[1]*dengxinyu@njust.edu.cn*, [2]*xuhuihui@njust.edu.cn*, [3]*lizihui@njust.edu.cn*,
[4]*baihuiwen@njust.edu.cn*, [5]*nilanfeng@njust.edu.cn*, [6]*zhangcz@njust.edu.cn*
Nanjing University of Science and Technology, 210094 (China)

Abstract

Research on gender differences has long been a prominent focus in academia. However, prior studies on gender disparities in research method selection have primarily concentrated on specific disciplines, lacking a comprehensive examination across the broader humanities and social sciences. This study addresses this gap by using Chinese humanities and social sciences as a case study, analyzing 63,742 PhD dissertations across 15 fields. After organizing the data and removing duplicates, 36 research methods were identified. A combination of large language models (LLMs) integration and manual retrieval was employed to detect the gender of PhD students and their supervisors. The findings reveal that 17 of the 36 research methods were strongly associated with the author's gender: male authors were more likely to employ summative theoretical construction methods, while female authors showed a preference for real-time data acquisition and analysis methods. Furthermore, significant differences were observed in the diversity of research methods used by supervisors based on gender, with female supervisors demonstrating a greater tendency toward methodological diversity. However, no significant relationship was found between the gender of PhD students and the diversity of research methods used. Dyadic analysis further highlighted that specific gender combinations significantly influenced preferences for particular research methods.

Introduction

In academic research, gender, as a critical variable, has attracted considerable attention due to the differences it generates (Bem, 1993; Eberhardt et al., 2023). Within the humanities and social sciences, gender plays a significant role throughout various stages of research, including project funding, paper publication, and domains such as the labor market, education, and politics. These gender differences are pervasive and have profound impacts (Allum, 2014; Ceci & Williams, 2011; Dolan, 2011). The selection and application of research methods, as a cornerstone of academic research, are also influenced by gender. Evidence indicates a correlation between the author's gender and their choice of research

methods (Diaz-Kope et al., 2019; Williams et al., 2018).

However, while the influence of researchers' gender on research method selection has been explored within specific disciplines, there remains a notable gap of research across the entire field of humanities and social sciences (Ashmos Plowman & Smith, 2011; Grant et al., 1987; Nunkoo et al., 2020). Examining gender differences in research method selection within these fields is essential for understanding academic research patterns and promoting the development of academic equity and diversity.

In gender-focused research, direct access to gender information of individuals is often unavailable, necessitating encoding based on names. Two empirically tested methods are widely used for this purpose: manual coding (Rajkó et al., 2023) and computational coding (Sebo, 2021b). While manual coding can be accurate, it is inefficient for large datasets, requiring significant time and resources, and limiting applicability and reliability. The emergence of large language models (LLMs) like ChatGPT provides a new avenue for gender inference, potentially addressing the shortcomings of traditional manual methods (Goyanes et al., 2024). However, existing limitations in LLMs, such as varying performance with different languages and uncommon names (Santamaría & Mihaljević, 2018), highlight the need for integrated approaches to improve gender inference accuracy. Strategies such as majority voting among multiple models and incorporating auxiliary information, such as institutional and disciplinary affiliations, have been proposed to enhance performance.

Given the lack of comprehensive research on gender differences in the selection of research methods among PhD students across the humanities and social sciences, and leveraging the capabilities of LLMs for gender inference, this study utilizes Chinese PhD dissertations as a corpus to explore gender differences in research method selection and their influencing factors. A combination of automated LLM-based inference and manual review was employed to improve the accuracy of gender detection for authors and supervisors. Statistical methods were then applied to analyze the relationship between gender and research method selection. This study aims to address the following research questions:

RQ1: What specific gender differences exist in the selection of research methods among PhD students and their supervisors in the humanities and social sciences?

RQ2: Does the gender of supervisors influence the gender differences in research method selection among PhD students, and if so, how does this influence manifest?

Related Work

In the humanities and social sciences, many scholars have explored gender differences from various angles. Although existing research has revealed gender differences in research topic selection, academic output and academic influence, there is still a lack of comprehensive research on gender differences in research method selection across the entire field. While some achievements have been made, limitations remain, and future research needs to delve deeper to fully understand the impact of gender on academic research method selection.

Gender differences in academia

In the humanities and social sciences, gender differences in academic research are an important topic, and many scholars have explored the manifestations of gender differences in academia from different perspectives. First, in terms of research topic selection tendencies, Kim et al.(2022) found that male and female scholars have distinct preferences in research topic selection. Female scholars tend to choose topics that focus on the rights of vulnerable groups and the coordination of social relationships, while male scholars prefer topics related to macro social structures, political systems, and economic development. Leahey(2006)found that female research projects are broader, spanning multiple subfields, while males tend to focus on fewer subfields, based on cumulative publications and unique keyword descriptors. Additionally, Zhang et al.(2021) pointed out that males focus more on scientific progress, while females pay more attention to social contributions, concluding that papers aimed at scientific progress have higher citation rates, while those aimed at social contributions have higher online reading rates. Second, in terms of gender differences in academic output, Male scholars submit more frequently to high-impact journals, while female scholars, being more cautious and setting higher standards, submit less often(Isabel et al., 2023).Specifically, female graduate students publish on average 8.5% fewer papers than male graduate students(Pezzoni et al., 2016). Among researchers of different age groups, the gender differences in research productivity also vary: among senior researchers, males generally have higher publication and citation counts than females, while among younger researchers, female participation has significantly increased, with publication and citation counts comparable to or even surpassing those of males, especially in high-impact research groups(Van Arensbergen et al., 2012). Finally, in terms of gender differences in the influence of academic achievements, the

"Matthew Effect" is evident in academic citations(Dion et al., 2018).Gender bias in citation practices is prevalent in multiple disciplines such as political science and economics(Ferber & Brün, 2011; Maliniak et al., 2013).Jayabalasingham(2020) found that although the overall average Field-Weighted Citation Impact (FWCI) ratio is close to 1, at the first-author level, males in most countries have a higher average FWCI than females.

Research methods in the humanities and social sciences

Research methods encompass the various means, techniques, and approaches used by scholars in the research process to explore and solve research problems scientifically, thereby obtaining reliable knowledge and conclusions(Trochim & Donnelly, 2001).

The classification of research methods in academic papers mainly includes manual classification and computer-automated classification(Chu & Ke, 2017; Eckle-Kohler et al., 2013). Early research primarily used manual classification to systematically sort, compare, and scientifically summarize the research methods used in different disciplines, such as sociology, library and information science, and management information systems, thereby constructing classification systems with universal applicability and disciplinary specificity(Chu & Ke, 2017; Palvia et al., 2003; Peritz, 1983). Manual classification relies on expert knowledge, ensuring high classification accuracy, and is suitable for research tasks requiring high classification accuracy and small data scales. However, manual classification is time-consuming and labor-intensive, and the scale of annotation is difficult to expand. As the number of research literature increases, manual classification struggles to meet the demands of large-scale data processing. With the development of machine learning technologies, computer-automated classification methods have gradually emerged. For example, in the field of Library and Information Science (LIS), Chu(2015) considered data collection and analysis techniques as the two core elements of research methods and classified research methods based on data collection techniques, dividing LIS research methods into 16 categories, including bibliometrics, content analysis, and the Delphi method. Zhang et al.(2021) developed rule-based methods for automatically identifying research methods in this field through content analysis and text mining. Zhang & Tian(2023) used deep learning models to automatically classify research methods in LIS.

Each discipline selects appropriate research methods based on its research characteristics and needs to better solve research problems within the discipline. Research has found that scholars in library and information science now use a

greater variety of research methods than before, with content analysis, experiments, and theoretical methods replacing the previously dominant survey and historical research methods(Chu, 2015).

Overview of research on method selection from a gender difference perspective

In academic research, the selection of research methods is one of the key factors influencing research outcomes, and whether gender affects research method selection has long been a topic of interest. Many scholars have explored this issue from different disciplinary perspectives, aiming to reveal the intrinsic relationship between gender and research method selection.

Grant et al. (1987) conducted a stratified random sampling study of 856 articles from 10 sociology journals between 1974 and 1983 and found that, regardless of the article's topic, female authors used qualitative methods more frequently than male authors. However, in articles related to gender topics, both male and female authors used quantitative methods more frequently than in non-gender articles. Dunn & Waller (2000)found males prefer secondary data and quantitative methods, while female authors were more likely to collect data through interviews and publish articles that did not include statistical analysis. Ashmos Plowman & Smith(2011) analyzed articles from four top management journals (1986-2008) and found female authors were significantly more represented in qualitative than non-qualitative research. Diaz-Kope et al.(2019)studied U.S. public affairs PhD programs and found that while males preferred quantitative methods and females leaned toward qualitative ones, females still chose quantitative methods more often. This indicates that the relationship between gender and research methods is not isolated but influenced by multiple external factors. Zhang et al. (2023) analyzed 5,281 articles from three top library and information science journals (1990-2019) and found significant gender differences in research methods. Specifically, female authors used interviews, surveys, and observations more frequently, while male authors preferred bibliometrics and theoretical methods. Thelwall et al. (2019) studied scholarly papers across various disciplines in the United States in 2017 and found that females were more likely to use exploratory and qualitative methods, while males preferred quantitative methods.

In summary, gender differences exist in academic research within the humanities and social sciences. Existing research on gender differences has demonstrated notable findings in areas such as topic selection, academic output, and influence. However, there is still a lack of comprehensive studies on gender differences in research method selection in the whole field. This paper will focus on this issue,

construct a theoretical framework, and promote academic fairness and diversity.

Data and Methodology

This paper uses PhD dissertations in the humanities and social sciences field as a corpus and combines LLMs with manual retrieval as the main research methods. The research framework of this paper is shown in Figure 1. First, journal paper databases and a corpus of over 60,000 PhD dissertations were collected, and gender matching was performed based on the "name + institution" rule. Then, LLMs were used to infer the gender of unmatched names, and the detection results were integrated, with manual retrieval for uncertain names. Finally, the gender information of PhD students and their supervisors was integrated, and correlation analysis was conducted to explore the relationship between gender and research method selection and its influencing factors.

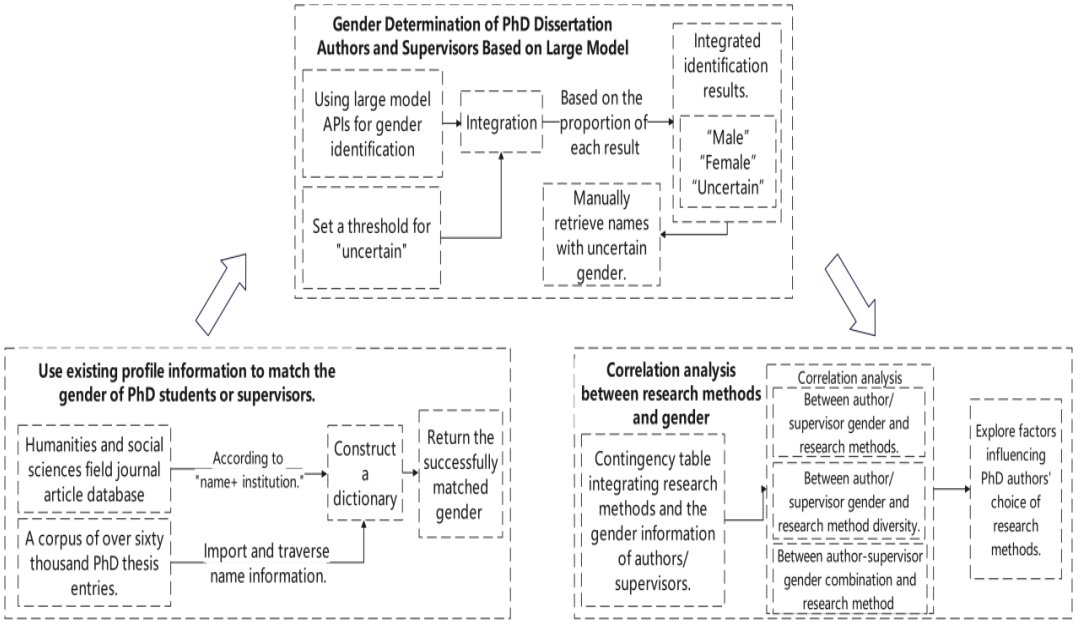


Figure 1. Framework of this study.

Dataset

The goal of this paper is to explore the manifestation of gender differences in research method selection and the underlying influencing factors by using LLMs to automatically infer the gender of authors and their supervisors in Chinese humanities and social sciences PhD dissertations. For this purpose, based on the

humanities and social sciences subject catalog established by other scholars, this paper systematically collected 63,741 PhD dissertation information from 1989 to 2020 from universities across the country. Each piece of information includes the dissertation title, publication year, author's name, author's primary discipline, author's institution, supervisor's name, and research methods extracted from the dissertation's research methods section. The research methods were recorded according to the classification framework of the humanities and social sciences field constructed by Zhang & Chu (2024). The number of samples in each discipline is shown in Table 1. After sorting and removing duplicates from the research methods of over 60,000 dissertations, 36 research methods were involved in 63,742 dissertations. The corpus used in this paper has a time span, sample size, and wide coverage of disciplines and regions, which improves the applicability of this research in different time periods and regions and its representativeness in the entire humanities and social sciences field.

Table 1. Primary Disciplines and Sample Sizes in the Chinese Humanities and Social Sciences.

Discipline	*Primary Discipline*	*#Secondary Disciplines*	*#Paper*
Philosophy	Philosophy	9	1998
Economics	Economics	20	11776
Law	Law	10	4305
Political Science	Political Science	9	2927
Sociology	Sociology	5	896
Ethnology	Ethnology	4	634
Marxism	Marxist Theory	6	3289
Educational Science	Educational Science	11	3080
Psychology	Psychology	4	704
Sports Science	Sports Science	5	1973
Literature	Chinese Language and Literature	9	5152
	Foreign language and Literature	9	1133
Journalism and Communication	Journalism and Communication	3	586

Artistic Discipline	Artistic Discipline	8	1305
Historical Science	Historical Science	11	3009
Management	Management Science and Engineering	1	7867
	Business Administration	7	6458
	Agroforestry Economic Management	2	3560
	Public Administration	6	2547
	Library and Information Science and Archival Management	6	542

Methodology

This paper addresses the gender detection of the collected corpus of over 60,000 PhD students and their supervisors, which lacks gender information. Since LLMs alone cannot guarantee high accuracy, gender information is first matched based on "name + institution" rule using existing databases to obtain some reliable gender information. Then, multiple platforms and eight LLMs are used for automatic gender detection and integration, with thresholds set to improve accuracy. Finally, manual retrieval is conducted for names with uncertain gender. At the same time, the correlation between gender and research methodology is analyzed by chi-square test and Mann-Whitney U test in terms of doctoral students, supervisors and their gender combinations to provide support for the subsequent research.

Gender Detection of Thesis Authors and Their Supervisors: This paper collected a corpus of over 60,000 records containing the names of PhD students and their supervisors, but the corpus only includes name information and lacks corresponding gender information, which needs to be supplemented. Since LLMs provide gender probabilities based on extensive data, the detection results have a certain degree of uncertainty. However, the gender obtained by matching the names of dissertation authors and their supervisors based on existing databases is highly reliable. Therefore, this paper first uses existing databases to match the gender of dissertation authors and their supervisors, and considers the gender of the matched part to be correct. The names of PhD students or supervisors that are successfully matched are filtered out from the corpus, and the remaining names are handed over to LLMs for automatic gender detection. This reduces the number of names for

which LLMs perform gender detection, indirectly improving the accuracy of the gender detection process. The gender detection steps are as follows:

Step 1. Gender matching of PhD students or supervisors using existing databases

This paper first uses PhD journal paper information databases for various disciplines to match the gender of the 60,000 PhD students and their supervisors in the corpus. The PhD journal paper information databases for various disciplines are formed by researchers screening PhD dissertation information from academic databases, extracting and organizing relevant content such as authors, instructors and titles for subsequent academic research analysis, including author and supervisor profile information, as well as "name-gender" columns for authors and supervisors.

This paper uses Python code to compare and match the PhD journal paper information databases with the corpus, thereby supplementing the gender information contained in the journal paper information databases into the name-only corpus used in this paper. The matching strategy is based on the rule that "name + institution" coincide simultaneously: first, a nested dictionary {name: {institution: gender}} is created from the journal paper information database; then, the author (supervisor) name information from the corpus is imported, and the return value after traversing the dictionary is the gender of the author(supervisor) at that institution. The matching success criteria is: if and only if the author's name exists and the institution in the corpus is included by the information of the institution in the journal paper, the gender result corresponding to that name will be returned. Ultimately, 10,153 author names were successfully matched, accounting for 15.09% of the total, and 37,516 supervisor names were successfully matched, accounting for 55.76% of the total, as shown in Table 2.

Since this step uses information from journal paper information databases, it is more convincing and credible, and can quickly and accurately obtain gender information for names.

Table 2. Name-Gender Information Matching for Authors and Supervisors.

Category	*#Successful Matches*	*Success Rate*
PhD Students	10,153	15.09%
Supervisors	37,516	55.76%

Step 2. Gender Inference and Integration of PhD Dissertation Authors and Supervisors Based on LLMs

Following the above steps of gender matching using the existing databases, the gender information obtained from the matching is filtered out. Then the remaining data without gender information is automatically inferred and integrated based on name information. This study innovatively adopts a LLMs integration approach for author gender detection, because of the particular challenges of Chinese name recognition and the limitations of existing methods. Existing studies have shown(Sebo, 2021a), that Chinese names lack explicit gender markers (e.g., western name suffixes), and the accuracy of traditional methods drops significantly when dealing with rare surnames or polyphonic characters. For example, the international platform Genderize.io recognizes Chinese pinyin with an accuracy of only 73.77% and is unable to deal with the ambiguity of polysyllabic characters. In contrast, LLMs learn the massive Chinese corpus such as ACL conference papers, and is able to infer names by combining their cultural background and regional characteristics, and reduces the uncertainty rate from 18.7% to 12.5% by the integration strategy(Zhang et al., 2023).

The automatic inference and integration are divided into two steps:

First, suitable tools are selected from numerous automatic gender detection tools. The "name2gender"[1] project in GitHub and the Genderize.io[2] platform were chosen. The "name2gender" project is a model for inferring gender from Chinese names based on LSTM. It utilizes the ccnc.csv and train.csv datasets containing names and corresponding genders. Data reading and preprocessing are handled by functions in the utils.py file. The model definition is in the name2gender.py file, including the Embedding layer, Dropout layer, LSTM layer, fully connected layer, ReLU activation function and Softmax layer. Model training is conducted using PyTorch's optimizer and loss function in finetune.py. Gender prediction for input names is performed in main.py.

For LLMs, priority was given to models provided by leading domestic internet companies. These companies are at the forefront of technological innovation and data processing, including Alibaba, Baidu, and 360. All of these companies are renowned for their strong technical capabilities and extensive industry influence. In addition to industry-leading LLMs, contributions from the academic field were also

[1] https://github.com/AlphaINF/name2gender.
[2] https://genderize.io/our-data.

considered. Consequently, models with academic backgrounds and research foundations, such as Qingyan and ChatGLM, were selected. These models not only have a good reputation in academia but also demonstrate excellent performance in specific field applications. Ultimately, it was decided to use two platforms and eight LLMs, including 360 Brain[3],ERNIE Bot[4], Baichuan[5], Qwen[6], Skywork[7], Qingyan[8], Doubao[9] and ChatGLM[10], for testing. APIs from each platform were invoked, and the previously matched 10,000+ PhD student author "name-gender" information was used as a detection sample, with 5,000 samples randomly selected for inference. The successfully matched "name-gender" results were used as reference answers to compare with the inference results of automatic gender detection tools.

The selection of automatic gender detection tools mainly considered two factors: inference accuracy and the manual retrieval proportion. In this research, the manual retrieval proportion is the percentage of names whose gender probabilities output by the model fall below the preset threshold and can't be automatically inference. The performance comparison results are shown in Table 3. After multiple automatic inferences and comparisons with correct gender information, it was found that the Genderize.io platform cannot recognize Chinese characters and is mainly used for recognizing Western names. It resulted in lower accuracy for Chinese names converted to pinyin. The gender inference threshold was set at 70%. After individual inference and integration, a comparison was made between the eight LLMs and the "name2gender" project. It was found that the eight LLMs showed higher accuracy and a lower proportion of manual intervention (the number of names with gender probabilities below 70% was relatively small). Consequently, the eight LLMs were chosen for inference and result integration in automatic gender detection, balancing accuracy and manual intervention.

[3] https://api.360.cn/v1/chat/completions.
[4] https://aip.baidubce.com/rpc/2.0/ai_custom/v1/wenxinworkshop/chat/completions_pro.
[5] https://api.baichuan-ai.com/v1/chat/completions.
[6] https://dashscope.aliyuncs.com/compatible-mode/v1.
[7] https://github.com/SkyworkAI/Skywork.
[8] https://open.bigmodel.cn/api/paas/v4/chat/completions.
[9] https://github.com/volcengine/volcengine-python-sdk.
[10] https://aip.baidubce.com/rpc/2.0/ai_custom/v1/wenxinworkshop/chat/chatglm2_6b_32k.

Table 1. Performance Comparison of Different Automatic Gender Detection Tools.

Platform/Model Name	*Accuracy*	*Proportion of Manual retrieval*
Genderize.io	73.77%	/
Name2gender	88.45%	14.3%
LLMs Integration	86.58%	12.5%

Second, after inferring the automatic gender detection tools, gender inference and integration of PhD dissertation authors and supervisors based on LLMs were conducted.

This study invoked the APIs of the eight LLMs, input the author and supervisor names from the dissertations, and the LLMs began the inference process. The inference results from the eight LLMs are integrated and returned, including the original name data and the gender ("male" or "female") and corresponding probability values inferred by each LLM.

To improve the accuracy of gender inference for PhD dissertation authors and supervisors in the humanities and social sciences, the inference results from the eight LLMs were integrated. In this process, a threshold of 0.7 was set to determine whether the gender is "uncertain". And the following integration strategy was used to determine the final gender inference result: first, the name was input into the eight LLMs for prediction; then, based on the proportion of uncertain predictions, i.e., the proportion of LLMs predicting "uncertain" to the total number of LLMs, the following steps were taken: (1) if the proportion of uncertain predictions was ≥50%, the gender cannot be determined; (2) if the proportion of uncertain predictions was <50%, the number of predictions for male and female was compared, and if the number of male predictions was greater than female, the gender is determined as male, and vice versa; (3) if the proportions of male and female predictions were both ≤50%, the gender also cannot be determined. This process, by integrating the prediction results of multiple models, aimed to improve the accuracy of gender inference.

After integration, the new results included all the content inferred by the LLMs, as well as the integrated results. The integrated results had three possibilities: "male," "female," and "uncertain," with 1,250 names having an integrated result of

"uncertain." This integration method leveraged the advantages of multiple LLMs, reducing single - model misjudgment risks and offering more reliable data for researching gender differences in research method selection.

Third, after the integration of LLMs, the genders corresponding to 5,076 PhD student names and 2,399 supervisor names could not be inferred. Considering the high cost of manual retrieval, the “name2gender” project in GitHub with high accuracy was utilized as an auxiliary tool to reassess the names with undetermined genders post-integration, thereby reducing the burden of manual retrieval. A consistent threshold of 70% was set as the boundary of whether to determine the gender. Following the secondary inference by GitHub, the genders corresponding to 861 author names and 405 supervisor name could not be inferred, so the remaining 1,266 names need manual retrieval.

Step 3. Manual Retrieval of PhD Dissertation Authors and Supervisors with "Uncertain Gender"

For the 1,266 names with "uncertain gender" in the integration results, further manual retrieval was needed to infer their gender. The following steps were taken to integrate the gender inference results from the LLMs.

First, the system filtered out names for which gender could not be inferred from the results obtained in Step 2. Then, the system searched for these filtered names within the corpus to retrieve corresponding name and institution information. Next, the system input the obtained name and institution information into a web browser for online retrieval. The retrieval process prioritized reliable sources, such as Baidu Baike[11] or the official websites of the author's or supervisor's institution, to obtain detailed information. In this paper, special attention was paid to the supervisors' personal homepage information during the search process. For entries with complete supervisor information, their institutional official websites or academic homepages were manually verified, and gender confirmation was performed through visual clues such as avatar photos. This supplementary verification mechanism enhances the credibility of supervisors' gender labelling, but fails to systematically address gender uncertainty in the PhD student population due to limited information on PhD students' networks. Afterward, the system verified the consistency of the name, gender, and institution information. Once verified, the system annotated the documents by accurately labeling the relevant information within the corpus, thus completing the retrieval process. Despite manual retrieval

[11] https://baike.baidu.com/

efforts, 576 names remained without corresponding gender information online. The failure to obtain results primarily occurred for PhD student authors, as individuals with low prominence or those who had left their institution were more difficult to identify. To ensure data accuracy, the system removed these 576 names and their corresponding dissertations from the corpus. These entries accounted for approximately 0.9% of the total number of dissertations, exerting minimal impact on the overall research.

Table 4. Summary of Gender Detection by Different Methods.

Method	*#Detection*	*#Certain Cases*	*#Uncertain Cases*
Gender Matching-Author	63741	9704	54037
Gender Matching-Supervisor	63741	35715	28026
LLMs Intergration-Author	54037	48961	5076
LLMs Intergration-Supervisor	28026	25627	2399
Name2gende-Author	5076	4215	861
Name2gende-Supervisor	2399	1994	405
Manual retrieval -Author	861	372	489
Manual retrieval -Supervisor	405	316	89

Table 4 presents the results of gender detection for each method. In this table, "Detection" indicates the number of name samples assigned to each method for gender identification. "Certain cases" denotes the number of name samples for which gender was successfully detected by each method. "Uncertain cases" represents the number of name samples for which gender could not be determined by each method.

Analysis of the Correlation Between PhD Students and Supervisor Gender and Research Methods: In order to investigate whether there is a correlation between the gender of PhD authors and their supervisors and the research methods they use in their dissertations, this paper analyzes the correlation between the gender of PhD students or their supervisors and the selection of research methods, the diversity of PhD students or their supervisors in terms of gender and the use of research methods, and the gender combination of PhD students and their supervisors and the selection of research methods from three perspectives.

Gender Differences of PhD Students or Supervisors and Research Method Selection: Based on the classification system of research methods in the field of humanities and social sciences, this paper organizes the gender information of PhD authors and their supervisors, and constructs a binary structure that demonstrates the relationship between gender and research methods selection.

In this structure, the gender variable is binary differentiated by 0 and 1, with 0 representing female and 1 representing male; the choice of research method is also represented by 0 and 1, with 0 indicating that a PhD student or supervisor did not use the research method in his/her dissertation, and 1 indicating that he/she did use the method.

In this paper, the chi-square test was used to analyze the relationship between the gender of PhD students or supervisors and the choice of research methods. The chi-square test is a statistical method applied to categorical data, and in this study, it was used to explore the relationship between two categorical variables, namely, the gender of PhD students or supervisors and the choice of research method. The $\chi 2$ value is obtained by dividing the square of the difference between the observed frequency and the expected frequency in each category, by the expected frequency, and then summing the results for all categories, with the specific formula as follows:

$$\chi^2 = \frac{(O_i - E_i)^2}{E_i}$$

Where, O_i denotes the number of theses in which PhD authors or supervisors of a certain gender used or did not use a particular research method in actual statistics; E_i represents the number of theses that used or did not use a particular research method under the assumption that there is no correlation between the gender of the PhD authors or supervisors and the choice of research method; By calculating the chi-square statistic and comparing the corresponding p-value and significance level, it is possible to determine whether there is a statistical correlation between the gender of PhD authors or supervisors and the choice of research methods.

Analysis of Gender Differences in Research Method Diversity Among PhD Students or Supervisors: Based on the dual structure between supervisor gender and research methods, data were secondary processed. Specifically, for each row of data labeled with supervisor gender, the number of research methods used (marked as 1) was summed to calculate the total number of research methods employed by

each supervisor. The same data processing approach was applied to the relationship between PhD authors' gender and the diversity of research methods.

For the significance analysis of the diversity of methods used by supervisors/authors, we use the Mann-Whitney U test. This non-parametric statistical test method is used to compare whether there is a significant difference in the medians of two independent samples, especially applicable to data that do not meet the normal distribution assumption(MacFarland & Yates, 2016).

Through the Mann-Whitney U test, it is possible to compare whether there is a significant difference in the median number of research methods between different gender groups. This method helps reveal the impact of gender on research method diversity, that is, to determine whether a certain gender tends to use more diverse research methods. Similarly, whether the asymptotic significance (two-tailed) p-value is less than 0.05 is used as the basis for determining whether gender is related to research method diversity, and the rank mean is observed to determine which gender of supervisor/PhD student uses more diverse research methods.

Analysis of the Correlation Between "PhD Student -Supervisor" Gender Combinations and Research Method Selection: Building upon the binary structure between the gender of PhD authors or supervisors and research methods, the data originally reflecting only the gender of PhD authors was expanded into a "PhD student - supervisor" gender combination format. This constructed a new binary structure to present the correspondence between the gender combination of “PhD students - supervisors” and the use of research methods. In this paper, the chi-square test is also used to analyze the significance of the relationship between the gender combination and the choice of research methods.

Results

This section will briefly describe the results of the analysis of the relationship between the gender of PhD students and their supervisors and the choice of research methods, and analyze the factors that affect the relationship between the gender of PhD students and their supervisors and the choice of research methods, in addition to drawing conclusions on how the gender factor affects the preference for and diversity of the choice of research methods.

Gender Differences Among PhD Students and Their Supervisors and Research Method Selection

This section answers RQ1. After conducting correlation analysis on the gender and research method selection of authors and supervisors in more than 60,000 PhD dissertations, this paper finds that PhD students or supervisors of different genders tend to choose specific research methods, and the gender of supervisors affects the diversity of research methods in dissertations, while the gender of authors does not have this effect.

Analysis of the Impact of Gender Differences on Research Method Selection: Among the 36 research methods, 17 methods showed strong correlations with author gender, and 15 methods showed strong correlations with supervisor gender. Males showed a clear preference for summative theoretical construction research methods, which may be related to long-standing societal expectations and cultivation of males in logical thinking and abstract construction. In educational and academic environments, males may be more encouraged to analyze problems from a macro, theoretical perspective, which is reflected in their research method selection. They are more adept at integrating existing knowledge systems and constructing systematic theoretical frameworks, so as to promote the deepening and expansion of disciplines at the theoretical level. Females showed a clear preference for real-time data acquisition and analysis methods, reflecting their high attention to actual situations and specific phenomena in the research process. They focus more on collecting first-hand information from the real world and uncovering patterns and trends through rigorous data analysis. This research method helps to bring academic research results that are closer to reality and more practically significant, supplementing and enriching the perspectives and content of academic research. For some new research methods, such as visual analysis and bibliometrics, the correlation with gender is not strong, but in the sample, female supervisors and female authors use them relatively more frequently.

For some research methods, the usage frequency is relatively low, which may not accurately reflect their actual application. Therefore, to ensure the reliability and persuasiveness of the research results, methods with usage instances below 1,000 were excluded from the correlation analysis. After screening and deleting these methods from the dataset, the correlation results shown in Figure 2 and Figure 3 were obtained.

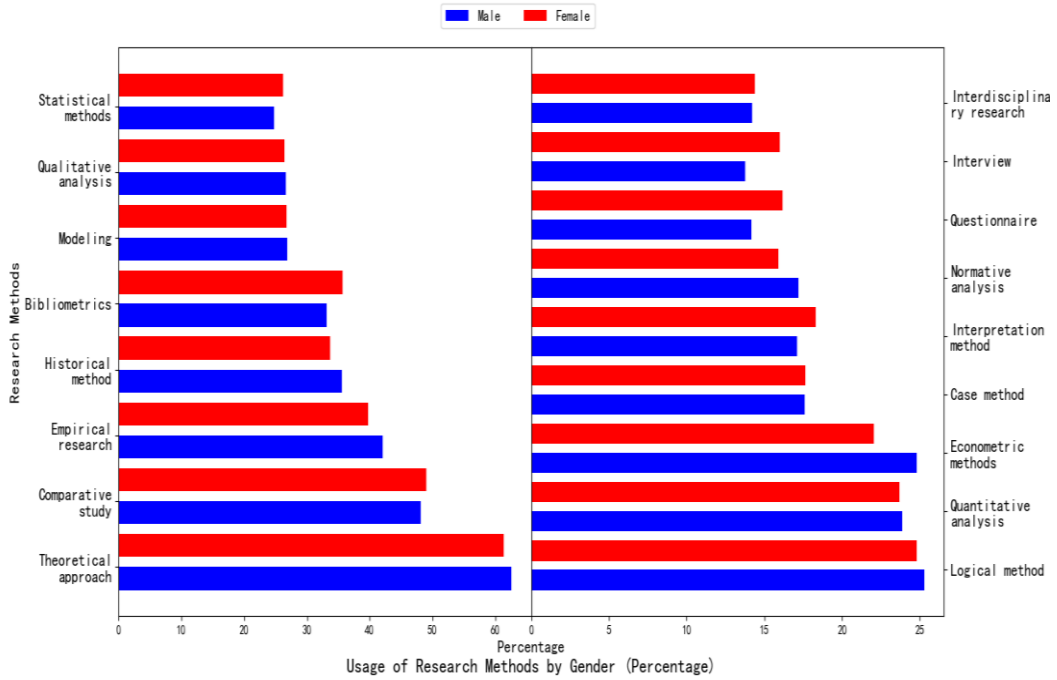


Figure 2. Results for Author Gender and Research Method Selection.

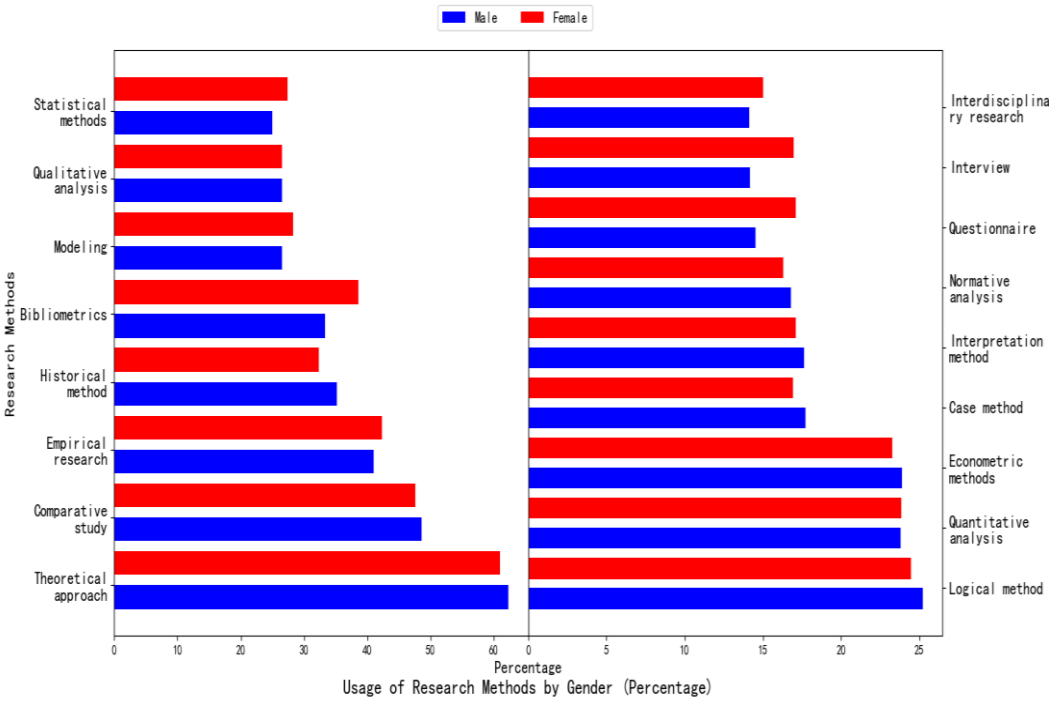


Figure 3. Results for Supervisor Gender and Research Method Selection.

Analysis of Gender and Research Method Diversity: The test results for supervisor gender and research method diversity and PhD student author gender and research method diversity are shown in Table 5. At the 5% significance level, the p-value for supervisor gender and research method diversity is 0.006 (<0.05), indicating a significant difference. Additionally, according to the rank mean values shown in Table 6, female supervisors (coded as 0) tend to use more diverse research methods. The calculated total entropy value for female supervisors is 3.49, slightly higher than that for male supervisors (3.48). This suggests that the distribution of method diversity among female scholars may be more uniform or complex. At the 5% significance level, the asymptotic significance (two-tailed) p-value for author gender and research method diversity is 0.515, greater than 0.05, indicating no significant difference between author gender and the diversity of research methods used. Additionally, as can be seen from the rank mean value in Table 6, the rank mean values of male and female authors are relatively close, further supporting the significant results.

Female supervisors tend to use more diverse research methods, reflecting their stronger inclusiveness and open-mindedness in the academic guidance process. They better recognize the strengths and applications of diverse methods, encouraging multi-perspective exploration, fostering academic innovation, and enhancing team creativity. In contrast, there is no significant difference in research method diversity among PhD students based on gender, which may mean that at the PhD student stage, gender has not yet had a decisive impact on the exploration of research method diversity. At this stage, PhD students are more influenced by disciplinary norms, the overall guidance style of supervisors, and the needs of the research topic itself, while individual gender factors play a relatively weaker role.

Table 5. Test Results of Gender and Diversity of Research Methods for Supervisors and Authors.

category	*The diversity of methods*			
	Mann - Whitney U	*Wilcoxon W*	*Z*	*Asymptotic Significance (Two - tailed)**
supervisor_sex	246275347.000	1696590500.000	-2.775	0.006
author_sex	462051071.000	1261591137.000	-0.651	0.515

Table 6. Distribution Statistics of Gender and Research Method Diversity.

category	*Distribution Variable*	*The diversity of methods*		
		N	*Rank mean value**	*Sum of ranks*
supervisor_sex	0	9312	32066.41	298602365.00
	1	53857	31501.76	1696590500.00
author_sex	0	23181	31646.68	733601728.00
	1	39988	31549.24	1261591137.00

Analysis of the Influence of Tutors on the Selection of Research Methods for PhD Students

This section answers RQ2. In the field of humanities and social sciences, tutor gender affects gender differences in PhD students' research method selection, either through stable transmission or gender interaction changing preferences.

Through the chi-square test, it was found that at the 5% significance level, 24 out of 36 research methods showed significant associations with gender combinations. For some research methods, the usage frequency is relatively low, which may not accurately reflect their actual application. Therefore, to ensure the reliability and persuasiveness of the research results, methods with usage frequencies below 0.07 were excluded from the correlation analysis. Filtered them out from the dataset, resulting in the correlation results shown in Figure 4. Additionally, in the two-dimensional analysis of "author-supervisor" gender combinations, methods that showed significant associations with gender combinations exhibited strengthened or weakened preferences for certain research methods in the same gender combinations. For example, the selection of theoretical analysis methods is related to gender, with male authors using them more frequently. The frequency of usage in gender combinations ranks as follows: male-male pairs (0.62) > male-female pairs (0.618) > female-male pairs (0.617) > female-female pairs (0.59).

The results of the two-dimensional analysis show that, to some extent, in a specific academic atmosphere or when the research team has formed a tradition of research method selection, the academic inheritance between supervisors and PhD students of the same gender is relatively stable. However, the combination of supervisors and PhD students of different genders reflects the influence of gender interaction on the choice of research methods. PhD students of different genders may collide with their supervisors and experience a different atmosphere of academic

exchanges, thus changing the preference of research methods, which is a phenomenon that provides rich research materials and certain inspiration for an in-depth understanding of the relationship between gender combinations in academic research.

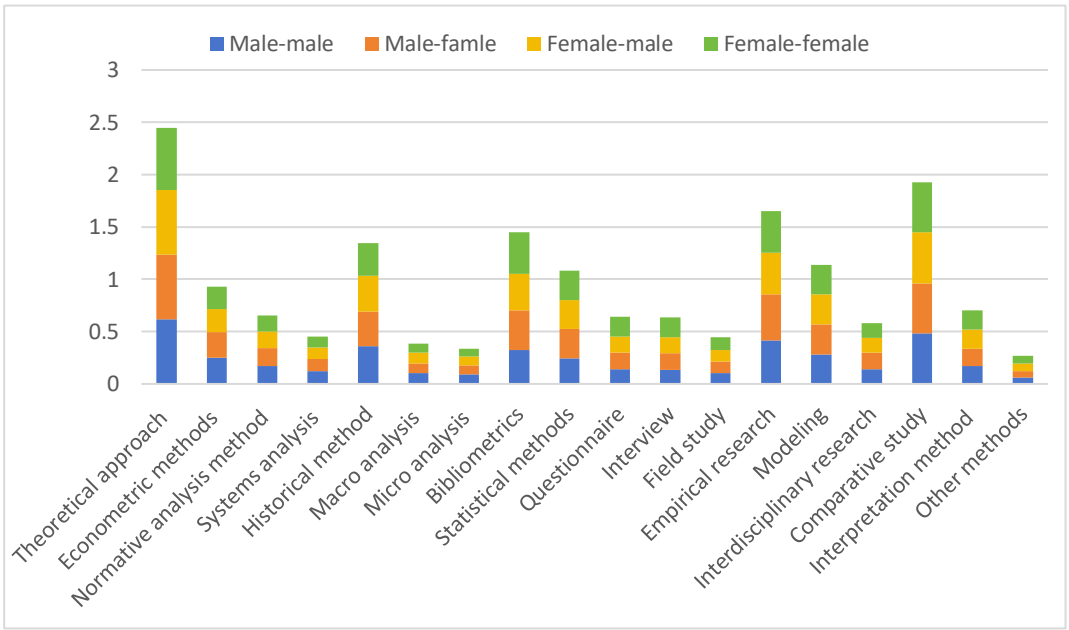


Figure 4. Correlation Results Between "Author-Supervisor" Gender Combinations and Research Method Selection.

Discussion

Research Implications

Theoretical Implications: This study has shown the impact of gender factors on the selection of academic research methods, providing a new perspective on how gender shapes academic research directions and methods. The study found that there are preference differences in research method selection among researchers of different genders. The difference may be related to gender roles in the socialization process, cognitive style differences, and gender stereotypes in academic environments. This paper also postulates that the diversity of research methods is influenced by the gender of the supervisor. Supervisor gender determines not only the composition but also the research culture and, as such, impacts research method adoption or diversity. These findings provide new theoretical dimensions to prior research on gender, emphasize the inclusion of gender factors in academic research,

and provide a theoretical basis for future gender difference research.

Practical Implications: Based on academic research and educational practice, this paper emphasizes the existence of gender differences and puts forward some suggestions. In academic research, supervisors should recognize the impact of gender differences in choosing research methods. Supervisors should encourage researchers to adopt diverse research methods to promote academic innovation and enrich knowledge production. In education, the educational institution should provide training and guidance to help students and researchers realize the existence of gender bias and encourage them to adopt a more fair and inclusive perspective in choosing research methods. It helps students and researchers be more aware of gender biases and develop a more impartial and inclusive approach when choosing research methods. This paper also puts forward that academic publishing units need to pay attention to the issue of gender bias in the review and editing process to ensure fairness and objectivity in the results of the research.

Research Limitations

Although this paper conducts a study of gender differences in the choice of research methods and the factors influencing them among PhD students in the whole field of humanities and social sciences, there are some limitations. First, the corpus, spanning from 1989 to 2000, may not fully capture contemporary research methodologies and trends. Second, while the study enhances the precision of gender detection through LLMs processing and manual retrieval, it cannot entirely eliminate discrepancies between automatically inferred gender and actual gender, leaving room for further improvement in the accuracy of research outcomes. Fourth, the study does not fully unpack the cultural mechanisms underpinning observed gender disparities. Additionally, generalisability is constrained by China's unique sociocultural and academic ecosystems. Institutional norms and cultural values likely interact to shape gendered method preferences. Future research should incorporate cross-regional comparisons, subdisciplinary analyses, and individual variables to explore cultural moderating effects.

Conclusion and Future Research Directions

This study focuses on gender differences in research method selection among PhD students and their supervisors in the Chinese humanities and social sciences. In terms of research method usage tendencies, among the various research methods examined, some showed significant gender correlations. Males more commonly

using theoretical construction methods and females more prominent in data collection and analysis methods. Regarding research method diversity, there are significant gender differences at the supervisor level. Female supervisors using more diverse research methods, with more complex and uniform method distributions than male supervisors. However, among PhD students, gender factors did not significantly affect research method diversity. In further dyadic analysis, a considerable number of research methods showed significant selection preference differences under different gender combinations, with some methods significantly increasing or decreasing in usage frequency under specific gender combinations However, this study has limitations in data coverage, the accuracy of gender detection, and the analysis of factors influencing gender differences in research method selection. Future research can comprehensively explore the influencing factors behind these differences from multiple perspectives such as research topics, the number of PhD students supervised by supervisors, and institutional levels.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Grant No.72074113).

References

Allum, J. (2014). Graduate Enrollment and Degrees: 2003 to 2013. *Washington, DC: Council of Graduate Schools*.

Ashmos Plowman, D., & Smith, A. D. (2011). The gendering of organizational research methods: Evidence of gender patterns in qualitative research. *Qualitative Research in Organizations and Management: An International Journal*, *6*(1), 64–82. https://doi.org/10.1108/17465641111129399

Bem, S. L. (1993). *The Lenses of Gender: Transforming the Debate on Sexual Inequality*. Yale University Press. http://www.jstor.org/stable/j.ctt1nq86n

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences*, *108*(8), 3157–3162. https://doi.org/10.1073/pnas.1014871108

Chu, H. (2015). Research methods in library and information science: A content analysis. *Library & Information Science Research*, *37*(1), 36–41. https://doi.org/10.1016/j.lisr.2014.09.003

Chu, H., & Ke, Q. (2017). Research methods: What's in the name? *Library & Information Science Research*, *39*(4), 284–294. https://doi.org/10.1016/j.lisr.2017.11.001

Dana Dunn, David V. Waller. (2000). THE METHODOLOGICAL INCLINATIONS OF GENDER SCHOLARSHIP IN MAINSTREAM SOCIOLOGY JOURNALS. *Sociological Spectrum*, *20*(2), 239–257. https://doi.org/10.1080/027321700279974

Diaz-Kope, L. M., Miller-Stevens, K., & Henley, T. J. (2019). An examination of dissertation research: The relationship between gender, methodological approach, and research design. *Journal of Public Affairs Education*, *25*(1), 93–114. https://doi.org/10.1080/15236803.2018.1463792

Dion, M. L., Sumner, J. L., & Mitchell, S. M. (2018). Gendered Citation Patterns across Political Science and Social Science Methodology Fields. *Political Analysis*, *26*(3), 312–327. https://doi.org/10.1017/pan.2018.12

Dolan, K. (2011). Do Women and Men Know Different Things? Measuring Gender Differences in Political Knowledge. *The Journal of Politics*, *73*(1), 97–107. https://doi.org/10.1017/S0022381610000897

Eberhardt, M., Facchini, G., & Rueda, V. (2023). *Gender Differences in Reference Letters: Evidence from the Economics Job Market*. https://doi.org/10.1093/ej/uead045

Eckle-Kohler, J., Nghiem, T., & Gurevych, I. (2013). Automatically assigning research methods to journal articles in the domain of social sciences. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1–8. https://doi.org/10.1002/meet.14505001049

Ferber, M. A., & Brün, M. (2011). The Gender Gap in Citations: Does It Persist? *The Gender Gap in Citations: Does It Persist?*, *17*(1), 151–158. https://doi.org/10.1080/13545701.2010.541857

Goyanes, M., de-Marcos, L., & Domínguez-Díaz, A. (2024). Automatic gender detection: A methodological procedure and recommendations to computationally infer the gender from names with ChatGPT and gender APIs. *Scientometrics*, *129*(11), 6867–6888. https://doi.org/10.1007/s11192-024-05149-2

Grant, L., Ward, K. B., & Rong, X. L. (1987). Is There An Association between Gender and Methods in Sociological Research? *American Sociological Review*, *52*(6), 856. https://doi.org/10.2307/2095839

Isabel, B., Ni, C., Badia, G., Tufenkji, N., Sugimoto, C. R., & Larivière, V. (2023). Gender differences in submission behavior exacerbate publication disparities in elite journals. *bioRxiv*. https://doi.org/10.1101/2023.08.21.554192

Jayabalasingham, B. (2020). *The researcher journey through a gender lens*.

Kim, L., Smith, D. S., Hofstra, B., & McFarland, D. A. (2022). Gendered knowledge in fields and academic careers. *Research Policy*, *51*(1), 104411. https://doi.org/10.1016/j.respol.2021.104411

Leahey, E. (2006). Gender Differences in Productivity: Research Specialization as a

Missing Link. *Gender & Society*, *20*(6), 754–780. https://doi.org/10.1177/0891243206293030

MacFarland, T. W., & Yates, J. M. (2016). *Introduction to Nonparametric Statistics for the Biological Sciences Using R*. Springer International Publishing. https://doi.org/10.1007/978-3-319-30634-6

Maliniak, D., Powers, R., & Walter, B. F. (2013). The Gender Citation Gap in International Relations. *International Organization*, *67*(4), 889–922. https://doi.org/10.1017/S0020818313000209

Nunkoo, R., Thelwall, M., Ladsawut, J., & Goolaup, S. (2020). Three decades of tourism scholarship: Gender, collaboration and research methods. *Tourism Management*, *78*, 104056. https://doi.org/10.1016/j.tourman.2019.104056

Palvia, P., Mao, E., Salam, A. F., & Soliman, K. S. (2003). Management Information Systems Research: What's There in a Methodology? *Communications of the Association for Information Systems*, *11*. https://doi.org/10.17705/1CAIS.01116

Peritz, B. C. (1983). Are methodological papers more cited than theoretical or empirical ones? The case of sociology. *Scientometrics*, *5*(4), 211–218. https://doi.org/10.1007/BF02019738

Pezzoni, M., Mairesse, J., Stephan, P., & Lane, J. (2016). Gender and the Publication Output of Graduate Students: A Case Study. *PLOS ONE*, *11*(1), e0145146. https://doi.org/10.1371/journal.pone.0145146

Rajkó, A., Herendy, C., Goyanes, M., & Demeter, M. (2023). The Matilda Effect in Communication Research: The Effects of Gender and Geography on Usage and Citations Across 11 Countries. *Communication Research*, 009365022211243. https://doi.org/10.1177/00936502221124389

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156. https://doi.org/10.7717/peerj-cs.156

Sebo, P. (2021a). How accurate are gender detection tools in predicting the gender for Chinese names? A study with 20,000 given names in Pinyin format. *Journal of the Medical Library Association*, *110*(2). https://doi.org/10.5195/jmla.2022.1289

Sebo, P. (2021b). Performance of gender detection tools: A comparative study of name-to-gender inference services. *Journal of the Medical Library Association*, *109*(3). https://doi.org/10.5195/jmla.2021.1185

Thelwall, M., Bailey, C., Tobin, C., & Bradshaw, N.-A. (2019). Gender differences in research areas, methods and topics: Can people and thing orientations explain the results? *Journal of Informetrics*, *13*(1), 149–169. https://doi.org/10.1016/j.joi.2018.12.002

Trochim, W. M., & Donnelly, J. P. (2001). Research Methods Knowledge Base. *Cincinnati, OH: Atomic Dog Publishing*.

Van Arensbergen, P., Van Der Weijden, I., & Van Den Besselaar, P. (2012). Gender differences in scientific productivity: A persisting phenomenon? *Scientometrics*, *93*(3), 857–868. https://doi.org/10.1007/s11192-012-0712-y

Williams, E. A., Kolek, E. A., Saunders, D. B., Remaly, A., & Wells, R. S. (2018). Mirror on the Field: Gender, Authorship, and Research Methods in Higher Education's Leading Journals. *The Journal of Higher Education*, *89*(1), 28–53. https://doi.org/10.1080/00221546.2017.1330599

Zhang C., & Chu X. (2024). Empirical Study on Application of Research Methods in Chinese Humanities and Social Sciences: A Large-Scale Investigation of PhD Dissertations(in Chinese). *Information studies: Theory & Application*, *47*(5), 48–57. https://doi.org/10.16353/j.cnki.1000-7490.2024.05.006

Zhang, C., & Tian, L. (2023). Non-synchronism in global usage of research methods in library and information science from 1990 to 2019. *Scientometrics*, *128*(7), 3981–4006. https://doi.org/10.1007/s11192-023-04740-3

Zhang, C., Wei, S., Zhao, Y., & Tian, L. (2023). Gender differences in research topic and method selection in library and information science: Perspectives from three top journals. *Library & Information Science Research*, *45*(3), 101255. https://doi.org/10.1016/j.lisr.2023.101255

Zhang C., Zeng J., Zhao Y. (2025). Is Higher Team Gender Diversity Correlated with Better Scientific Impact? Journal of Informetrics, 19(2): 101662. https://doi.org/10.1016/j.joi.2018.12.002

Zhang, L., Sivertsen, G., Du, H., Huang, Y., & Glänzel, W. (2021). Gender differences in the aims and impacts of research. *Scientometrics*, *126*(11), 8861–8886. https://doi.org/10.1007/s11192-021-04171-y

Zhang, Z., Tam, W., & Cox, A. (2021). Towards automated analysis of research methods in library and information science. *Quantitative Science Studies*, *2*(2), 698–732. https://doi.org/10.1162/qss_a_00123

https://doi.org/10.51408/issi2025_012

Gender Disparities in Academic Research: A Comparative Study of Armenia and Italy

Shushanik Sargsyan[1], Edita Gzoyan[2], Giovanni Abramo[3], Ciriaco Andrea D'Angelo[4]

[1]*shushaniksargsyan8@gmail.com*
Institute for Informatics and Automation Problems of NAS RA
1 Paruyr Sevak St, 0014, Yerevan (Republic of Armenia)

[2]*editagzoyan@gmail.com*
Institute for Informatics and Automation Problems of NAS RA
1 Paruyr Sevak St, 0014, Yerevan (Republic of Armenia)

[3]*giovanni.abramo@unimercatorum.it*
Universitas Mercatorum, Laboratory for Studies in Research Evaluation
Piazza Mattei 10, 00186 Rome (Italy)

[4]*dangelo@dii.uniroma2.it*
University of Rome "Tor Vergata", Dept of Business Engineering
Via del Politecnico 1, 00133 Rome (Italy)

Abstract

Gender disparities in academic research are a critical concern in the quest for equality in science and higher education. These disparities are evident in research output, citation impact, collaboration networks, and representation in senior academic roles, with women generally underrepresented and displaying lower performance metrics compared to men. However, the nature and extent of these gaps often differ across countries due to varying cultural and institutional contexts. This study examines gender differences in research performance in STEMM fields by comparing Armenia and Italy, two nations with distinct academic traditions and gender norms. Using 2017–2021 data from the Web of Science core collection, the proposed analysis encompasses over 3,600 Armenian and 27,000 Italian scientists, evaluating metrics such as publication counts, citation impact, and collaboration patterns at the individual level. The findings highlight how national contexts shape the gender gap in research performance, revealing unique barriers faced by female researchers in each setting. By investigating these disparities through a comparative lens, the study provides insights into the complex interplay between gender and geography in academic research. These insights aim to inform policy measures tailored to address gender-based inequities in diverse academic environments.

Introduction

Gender disparities in research performance and academic career advancement have become central issues in the discourse on equality in science and higher education (Larivière et al., 2013; Elsevier, 2020). These disparities manifest in various forms, including differences in research output, citation impact, collaboration networks, and representation in senior academic positions (Ceci & Williams, 2011; Bendels et al., 2018). While the general pattern of underrepresentation of women and lower research performance metrics compared to their male counterparts is well documented, the degree and nature of these disparities often vary significantly across countries and cultural contexts (UNESCO, 2019). Understanding these differences

is crucial for developing policies that address the unique barriers faced by female researchers, particularly in contexts where academic and research traditions vary widely (Huang et al., 2020).

This study provides a comparative analysis of gender differences in research performance between Armenia and Italy, two countries with distinct historical, cultural, and institutional backgrounds that shape academic norms and gender roles in different ways. Armenia, a post-Soviet country in the Caucasus region, is undergoing rapid socio-economic development, including increased attention to gender equality (Yeritsyan, 2019). However, Armenia still faces considerable challenges related to traditional gender roles, particularly in high-skill, male-dominated sectors (UNDP Armenia, 2020). In academia, the barriers faced by female researchers can be exacerbated by structural limitations in research funding, limited networking opportunities, and insufficient institutional support, which can impact their research performance and visibility in the academic community (van den Besselaar & Sandström, 2016).

In contrast, Italy is a Western European country with a well-established higher education system and more progressive gender equality policies, especially within academia (Bettio & Verashchagina, 2009). Despite this, Italy's academic sector exhibits a notable gender gap in terms of senior leadership positions, publication metrics, and research funding opportunities, particularly in fields like engineering and the physical sciences (Moscatelli et al., 2019). Italian female researchers often confront institutionalized biases and slower career progression, particularly as they approach senior academic ranks, contributing to gendered differences in research productivity and impact (Guarino & Borden, 2017; Mairesse & Pezzoni, 2015). Comparing Armenia and Italy thus allows one to analyze how gender disparities in research performance manifest across contrasting socio-cultural and academic environments (Abramo, Aksnes, & D'Angelo, 2021; Addis & Villa, 2003).

Research performance can be analyzed through a combination of quantitative indicators, including publication counts, citation impact, and collaboration patterns. These metrics provide insight into the scholarly productivity, influence, and networking capabilities of researchers and reveal potential barriers specific to gender (Bozeman & Corley, 2004). For instance, prior research has indicated that female researchers, on average, tend to have lower publication rates and citation impacts than male researchers, potentially due to unequal access to resources, disproportionate administrative and teaching responsibilities, and biases in peer review processes (Abramo, D'Angelo, & Rosati, 2016; Dworkin et al., 2020; Witteman et al., 2019). Additionally, gender differences in collaboration networks can influence access to co-authorship opportunities and interdisciplinary partnerships, both of which are critical for academic success and impact (Abramo, D'Angelo, & Di Costa, 2019; Caplar, Tacchella, & Birrer, 2017; Thelwall & Wilson, 2014).

The primary aim of this study is to compare gender differences in research performance between Armenia and Italy, focusing on three core aspects: (1) publication output, (2) citation impact, and (3) productivity. By analyzing these metrics across gender lines, this study seeks to identify the extent to which the gender

gap in research performance is influenced by the national context and to explore the underlying factors contributing to these differences. In doing so, it offers a nuanced understanding of how gender and geographical context interact to shape research performance (Aksnes, Rorstad, & Sivertsen, 2011).

Furthermore, this comparative study seeks to inform policymakers and academic institutions in Armenia, Italy, and beyond about potential interventions to promote gender equity in academia. For instance, differences in citation impact could indicate the need for policies that reduce barriers to accessing high-impact journals and conferences (Elsevier, 2020; Stoet & Geary, 2018). Ultimately, this research contributes to the broader goal of creating equitable academic environments where researchers of all genders can achieve their full potential.

Literature review

Gender disparities in academia have been widely documented across multiple dimensions, including research productivity, career advancement, and leadership positions. A growing body of research shows that female researchers often publish fewer papers than their male counterparts, achieve fewer citations, and have less access to collaborative networks, which collectively impact their academic influence and visibility (Larivière et al., 2013; Bendels et al., 2018). These disparities are typically attributed to a combination of structural, institutional, and cultural factors that hinder women's academic progression, such as unequal distribution of research funding, higher teaching or service burdens, and biases in publication and peer review processes (Ceci & Williams, 2011; Witteman et al., 2019).

Cross-national studies have increasingly highlighted that gender disparities in research performance are not universal but instead vary significantly by country, discipline, and institutional framework (Elsevier, 2020). For instance, countries with robust gender equality policies, such as those in Northern Europe, often exhibit smaller gender gaps in research output and impact compared to countries where such policies are less established (UNESCO, 2019).

This showed also during the COVID-19 pandemic. Differently from common belief, only in the Far East, women experienced a worse decrease in research output with respect to men. In the U.S. and China female and male scholars reduced their research output at a similar rate. In Europe, contrasting evidence emerged. In some countries (France, Netherlands and Switzerland) women were hurt more than men; in others (Germany and Spain) the opposite holds true, while in such countries as Italy, Sweden and U.K. gender differences are hardly noticeable (Abramo, D'Angelo, & Mele, 2022).

These variations emphasize the importance of contextual factors in shaping academic gender disparities and underscores the need for comparative studies to deepen our understanding of how different socio-cultural and institutional contexts contribute to these disparities.

Research output, typically measured by the number of publications, remains a key metric for academic success and is often influenced by gender. Studies consistently show that, on average, female academics publish fewer papers than their male colleagues, a disparity that has been observed across disciplines, including STEM

fields and social sciences (Aksnes, Rorstad, & Sivertsen, 2011; Huang et al., 2020). Various factors contribute to this gap, including differences in time allocation between research and other responsibilities such as teaching and administration, which often fall disproportionately on women (Guarino & Borden, 2017). Additionally, women in academia may face greater challenges in securing research funding, which directly affects their ability to conduct and publish high-quality research (van den Besselaar & Sandström, 2016).

Notably, recent research has examined the "leaky pipeline" phenomenon, wherein female representation in academia decreases at each successive career stage, especially in higher academic ranks (Alper & Gibbons, 2017). This effect is often pronounced in countries with traditional gender roles, where female academics may face greater cultural expectations around caregiving responsibilities, thereby limiting their time for research and collaboration. Armenia and Italy both experience significant "pipeline leakage," particularly in senior positions, though the underlying causes and extent of this trend differ between the two countries (Greska, 2023; Borrell-Damián & Rahier, 2019).

Citation-based metrics are widely used to assess research impact and visibility in the academic community. Studies show that female researchers generally receive fewer citations than male researchers, even after controlling for publication volume and field-specific citation rates (Dworkin et al., 2020). This disparity has been attributed to a range of factors, including potential biases in citation practices and the gendered dynamics of academic networks, which can affect the visibility and perceived impact of women's research (Caplar, Tacchella, & Birrer, 2017).

Gender differences in citation impact are also influenced by the nature of the journals where female researchers publish. Women are often underrepresented in high-impact journals and may experience greater difficulty in accessing these prestigious publication venues due to biases in the editorial process or fewer collaborative opportunities that lead to impactful research outputs (Addis & Villa, 2003). The Armenian context, where academic journal publishing is still developing, poses additional challenges for researchers, particularly for women, who may have limited access to international platforms with high visibility. In contrast, Italian researchers benefit from more established networks and access to high-impact publication venues, though significant gender gaps persist, particularly in STEMM disciplines[1], and among top scientists (Abramo, D'Angelo, & Caprasecca, 2009a; Abramo, D'Angelo, & Caprasecca, 2009b).

Collaboration is an increasingly vital component of academic success, as researchers who collaborate extensively tend to publish more and achieve higher citation rates (Mairesse & Pezzoni, 2015). However, studies indicate that women are often less integrated into influential academic networks and may have fewer opportunities for international and interdisciplinary collaboration (Bozeman & Corley, 2004). Limited access to collaboration networks can hinder women's research output and impact, contributing to the observed gender disparities in academic performance.

[1] Science, technology, engineering, mathematics, and medicine.

Research by Thelwall and Wilson (2014) suggests that women are more likely to collaborate within their institutions and less likely to engage in international collaborations, which tend to be more productive and impactful. This trend is especially relevant in countries with limited research infrastructure, such as Armenia, where collaborative opportunities with international peers may be constrained by institutional and funding limitations. In Italy, where the academic landscape is more globally integrated, female researchers face fewer structural barriers to international collaboration but still encounter challenges in forming and sustaining partnerships in male-dominated fields (Abramo, D'Angelo, & Murgia, 2013;).

The academic gender gap in Armenia reflects broader societal dynamics, as Armenia's recent post-Soviet transition has influenced both its educational infrastructure and gender norms in professional settings. Traditional gender expectations, combined with limited institutional support for women in research, contribute to gender disparities in research performance (Yeritsyan, 2019). Armenia's nascent efforts to address gender equality have yet to overcome these entrenched norms fully, and female researchers may experience significant structural and cultural barriers to academic success (UNDP Armenia, 2020).

Italy, on the other hand, is a Western European country where gender equality in academia has been progressively recognized and addressed through various policies. However, Italy's academic sector still reflects significant gender biases, especially in senior academic roles. Female representation decreases sharply in higher academic ranks, and Italian women researchers in science and engineering fields encounter particularly strong barriers to promotion and access to research funding (Bettio & Verashchagina, 2009; Moscatelli et al., 2019). Additionally, family-oriented cultural expectations in Italy often result in career interruptions for female researchers, which can negatively affect their research output and overall academic impact.

The findings from these studies underscore the importance of targeted policy interventions to address gender disparities in academia. Research suggests that policies that provide flexible career paths, support family-friendly work environments, and promote equitable access to research funding can reduce gender gaps in research output and impact (Stoet & Geary, 2018). Moreover, initiatives aimed at enhancing collaborative opportunities and mentorship programs can support female researchers in building stronger academic networks, thereby improving their access to high-impact publication channels and collaborative research opportunities.

In Armenia, policy efforts focused on building a more inclusive research environment and increasing access to international networks may benefit female researchers by alleviating structural limitations. For Italy, addressing gender disparities in senior academic roles and ensuring transparency in promotion and funding processes could promote gender equity at higher academic levels.

Data and methods

The census of Armenian scientists and their publication portfolio

We carried out the census of the research staff of the Armenian national science system, collecting names of professors and researchers: i) from the official websites of higher education institutions and research centres of the National Academy of Science of the Republic of Armenia (NAS RA); ii) sending official letters to the respective organizations with the request to provide the necessary information; and iii) harvesting the necessary information from the financing agreements of the research institutions of NAS RA, available on the web page of the Government.
Overall, we obtained microdata for 20 research organizations of the NAS RA and 14 universities involved in STEMM research, i.e. personal identifiers, affiliations, full names, gender, and academic rank. At the next stage we collected publications from Web of Science, having "Armenia" as affiliation country, and manually matching: i) the researchers' full names previously obtained with the author list; ii) the official affiliation with the bibliometric address list. Finally, we measured precision and recall of our bibliometric dataset, by manually checking data on a random sample.

The census of Italian scientists and their publication portfolio

The MUR maintains a database of university personnel. For each professor, this database provides information on their name and surname, gender, affiliation, discipline classification, and academic rank at the close of each year.[2] A similar database does not exist for public research institutions, which forces us to restrict the Italian census to professors only. For reasons of significance, our analysis is limited to those professors who held formal faculty positions for at least three years over the 2017-2021 period. The bibliometric dataset used to assess professors' output is extracted from the Italian Observatory of Public Research (ORP), a database developed and maintained by Abramo and D'Angelo and derived under license from the Clarivate Analytics Web of Science (WoS) Core Collection. Beginning from the raw data of the WoS, we first reconcile the author's affiliations, and then apply a complex algorithm to disambiguate the true identity of the authors. In ORP each publication is attributed to the university professors that produced it.[3]

Standardizing academic rank and classifying researchers by field

Since the dataset for Italy includes exclusively university professors, we will also use the term "professor" for all Armenian individuals. For this purpose, the ranks of the research staff of NAS RA institutions were matched to the equivalent academic rank as follows: Research director => Full professor; Senior researcher => Associate professor; Researcher => Assistant professor.
For benchmarking the two national systems, it is key to categorize each professor in the dataset into a specific scientific discipline. To achieve this, we utilized the WoS

[2] http://cercauniversita.cineca.it/php5/docenti/cerca.php, last accessed on 1 July 2024.
[3] The harmonic average of precision and recall (F-measure) of authorships, as disambiguated by the algorithm, is around 97% (2% margin of error, 98% confidence interval).

classification scheme and: 1) identified the WoS indexed publications of each professor under observation; 2) assigned to each publication the SC or SCs of the hosting journal; 3) classified each professor in the most recurrent SC in their publication portfolio.
A problem arises when the portfolio is limited to one or a few publications or when one observes more than one dominant SC. At this purpose, such analysis was carried out on an extended time window of eleven years (2010-2022). Residual cases of professors with more than one dominant SC were solved by randomly selecting one of the dominant SCs.

The final dataset

Because of the limited coverage of publications in the Arts and Humanities, for reasons of significance, we included in the analyses only professors in STEMM SCs (Larivière, Archambault, Gingras, Vignola-Gagné, 2006; Aksnes & Sivertsen, 2019). Moreover, after merging the datasets of the two countries, we included in the final dataset only those SCs (128 in all) with at least one Armenian and one Italian professor. The final dataset consists of 3617 Armenian and 27034 Italian professors. Their distribution per field[4] is shown in Table 1.

Table 1. Dataset of analysis.

Field	No. of SCs	No. of Armenian professors	No. of Italian professors
Biology	26	637 (17.6%)	5232 (19.4%)
Biomedical Research	12	387 (10.7%)	2863 (10.6%)
Chemistry	8	332 (9.2%)	1566 (5.8%)
Clinical Medicine	25	437 (12.1%)	5425 (20.1%)
Earth and Space Sciences	11	309 (8.5%)	2271 (8.4%)
Engineering	27	708 (19.6%)	5487 (20.3%)
Mathematics	3	250 (6.9%)	1496 (5.5%)
Physics	16	557 (15.4%)	2694 (10.0%)
Overall	128	3617	27034

In both countries, Engineering is the most represented field while Mathematics is the one with the fewest number of professors on staff. While the distribution by field is relatively similar between the two countries, the breakdown in the three academic ranks is very different. Full professors in the Italian dataset account for 31% of the total, compared to 12.5% for Armenia. In contrast, Italian assistant professors are 16.7% of the total, while for Armenia they are almost 60%.

[4] SCs are grouped in fields following a pattern previously published on the website of ISI Journal Citation Reports, but no longer available on the current Clarivate portal. There are no cases in which an SC is assigned to more than one field.

Measuring research performance

The comparative evaluation of the research performance of individual professors is proxied by an output-to-input productivity indicator named Fractional Scientific Strength (FSS),[5] defined as:

$$FSS_p = \frac{1}{\left(\frac{w}{2} + k\right)} \cdot \frac{1}{t} \sum_{i=1}^{N} c_i \ f_i \qquad [1]$$

where:

w = average yearly salary of the professor (we halve labor costs, assuming that 50 percent of professors' time is allocated to activities other than research);

k = average yearly capital available for research to the professor;

t = number of years of work by the professor in the period under observation;

N = number of publications by the professor in the period under observation;

c_i = impact of publication i (weighted average of the discipline-normalized citations received by publication i and the discipline-normalized impact factor of the hosting journal);[6]

f_i = fractional contribution of professor to publication i;[7]

As for the input factors (w and k), we relied on Abramo, Aksnes, & D'Angelo (2020, Table 4).

For each professor,[8] FSS is computed in absolute value and percentile rank, by comparison with the same data referring to all professors in the same subject category in the dataset.

The analysis will also be conducted through indicators that measure the different components of FSS and, more specifically, the output (O_p), the fractional output

[5] For a comprehensive explanation of the methodology, underlying theory, assumptions and limitations, as well as the input data source, we direct the reader to Abramo and D'Angelo (2014) and Abramo et al. (2020).

[6] This combination serves as the most accurate projection of future long-term citations for a publication (Abramo et al., 2019). Citations are adjusted to the mean of the distribution concerning all referenced publications from the same year and the Web of Science subject category (SC) of publication i. The journal's impact factor (IF), corresponding to the year of publication, is normalized relative to the average of the IF distribution of all journals in the same SC of publication i.

[7] In the field of life sciences in Italy, it is customary for authors to delineate their respective contributions to published research based on the order of names in the byline. In SCs related to these areas, we assign varying weights to each co-author depending on their position in the byline and the nature of the co-authorship (intra-mural or extra-mural). When the first and last authors are affiliated with the same university, each is attributed 40% of the citations, with the remaining 20% distributed among all other authors. If the first two and last two authors come from different universities, 30% of citations go to the first and last authors, 15% to the second and penultimate authors, and the remaining 10% is divided among all other contributors. These weighting values were determined with guidance from eminent Italian life sciences scholars and can be adjusted to align with various practices in other national contexts. In all other subject areas, fractional contribution is calculated as the inverse of the number of authors.

[8] As for the research staff of Armenian researchers working at NASRA institutes, we equate the research unit leader to full professor, senior researcher to associate professor, and researcher to assistant professor.

(FO_p) and average impact, as measured by standardized citations (AI_p), and hosting journals' standardized impact factors (JI_p). For this purpose, we will use the indicators described below.

$$O_p = \frac{N}{t} \quad [2]$$

$$FO_p = \frac{1}{t}\sum_{i=1}^{N} f_i \quad [3]$$

$$AI_p = \frac{1}{N}\sum_{i=1}^{N} cit_i \quad [4]$$

$$JI_p = \frac{1}{N}\sum_{i=1}^{N} if_i \quad [5]$$

With
N = number of publications by the professor in the period under observation;
f_i = fractional contribution of professor to publication i;
cit_i = year- and discipline-normalized citations received by publication i;
if_i = discipline-normalized impact factor of the hosting journal at the year of publication.

Results

The incidence of women in the research staff of the two countries

Table 2 and Figure 1 provide a comparative view of the gender distribution within the research staff of Armenia and Italy, categorized by academic rank and field. Notably, Armenia exhibits a higher overall representation of women in STEMM fields (52.0%) compared to Italy (35.8%). This trend is particularly evident in Biology and Biomedical Research, where the share of female researchers in Armenia exceeds 70%, whereas Italy reports approximately 50% female participation. Conversely, in male-dominated fields like Engineering and Physics, both countries show significantly lower female representation. In these fields, only 42.9% of Armenian researchers and 21.8% of Italian researchers are women in Engineering, and 28.9% (Armenia) and 19.0% (Italy) in Physics.

Interestingly, the concentration index (CI) reveals that women in Armenia are more proportionally represented across various fields compared to the national average, while Italy shows more pronounced disparities in female representation across disciplines. These data suggest systemic differences in how gender roles manifest in academic environments, with Armenian women achieving higher numerical participation but potentially facing other structural barriers.

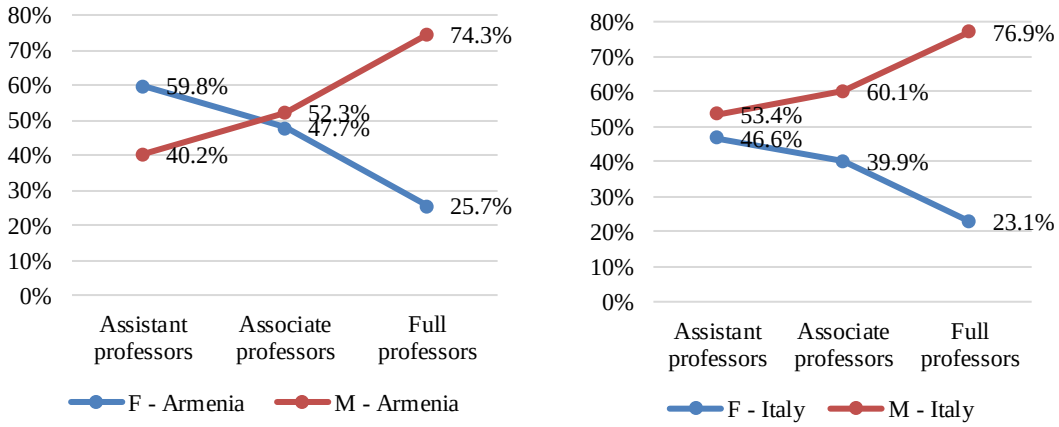


Figure 1. Research staff of the two countries in the dataset, by gender and academic rank.

Table 2. Share of female professors, by gender, country and field.

	Armenia		Italy	
Field	Share of females	CI*	Share of females	CI*
Biology	70.6%	1.358	49.6%	1.385
Biomedical Research	71.3%	1.371	48.4%	1.351
Chemistry	60.2%	1.158	44.7%	1.248
Clinical Medicine	60.9%	1.170	35.9%	1.003
Earth and Space Sciences	47.9%	0.921	35.1%	0.981
Engineering	42.9%	0.825	21.8%	0.608
Mathematics	30.8%	0.592	36.7%	1.024
Physics	28.9%	0.556	19.0%	0.531
Overall	52.0%		35.8%	

** concentration index, given by the share of female professors of a country in a given field divided by the share of female professors of that country overall. A value of 1.2 means that in the field, females are 20% more than their expected value measured at the overall country level.*

Output

Table 3 details the percentage of professors with at least one WoS publication during the 2017–2021 period. The data reveal stark contrasts in research output between Armenia and Italy. In Italy, the vast majority of professors (98.1%) have at least one WoS publication, with negligible gender differences across fields. In Armenia, however, the overall share is markedly lower at 28.3%, with significant variations by field and gender. For instance, while 39.1% of Armenian women in Physics have at least one publication, the percentage drops to just 15.6% in Mathematics. This pattern highlights not only a productivity gap between the two countries but also variations within Armenia that suggest field-specific challenges for female researchers.

The combination of financial limitations, lack of integration into international networks, language barriers, institutional publication practices, and broader societal inequalities likely explains the disparities in research output between Armenia and Italy.

Figures 2 and 3 further explore the Armenian context, showing how affiliation type and the number of affiliations correlate with publication activity. Women with multiple affiliations tend to exhibit higher publication rates, hinting at the potential role of collaborative opportunities in mitigating structural barriers to research output. Figure 4 underscores the disparity in publication activity by academic rank, with full professors in both countries demonstrating the highest productivity rates. However, the gender gap persists, particularly at senior levels in Armenia, suggesting entrenched structural challenges.

Table 3. Share of professors with at least one 2017-2021 WoS publication, by gender, country, and field.

	Armenia			Italy		
Field	F	M	Total	F	M	Total
Biology	22.2%	22.5%	22.3%	99.0%	98.9%	99.0%
Biomedical Research	27.2%	26.1%	26.9%	99.3%	98.9%	99.1%
Chemistry	33.0%	43.9%	37.3%	99.4%	99.2%	99.3%
Clinical Medicine	27.4%	31.0%	28.8%	98.1%	98.0%	98.0%
Earth and Space Sciences	19.6%	25.5%	22.7%	97.5%	97.4%	97.4%
Engineering	15.1%	18.6%	17.1%	98.0%	98.1%	98.1%
Mathematics	15.6%	35.8%	29.6%	93.3%	95.8%	94.9%
Physics	39.1%	50.8%	47.4%	95.7%	97.3%	97.0%
Overall	24.7%	32.3%	28.3%	98.1%	98.0%	98.1%

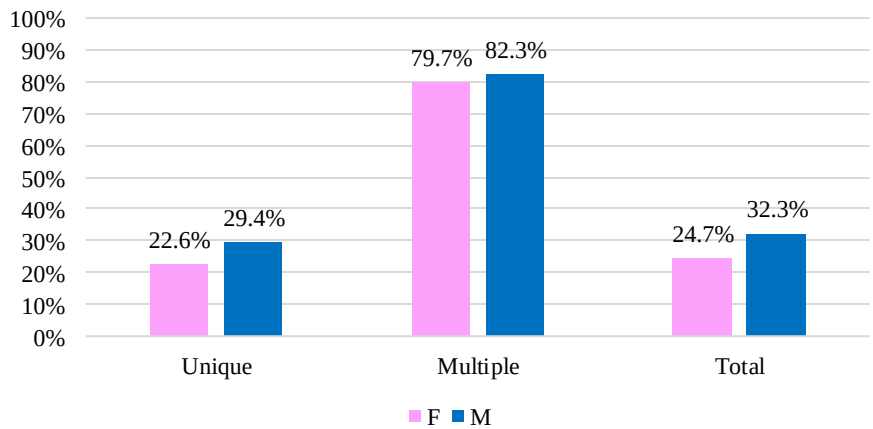


Figure 2. Share of Armenian professors with at least one 2017-2021 WoS publication, by gender and number of affiliations.

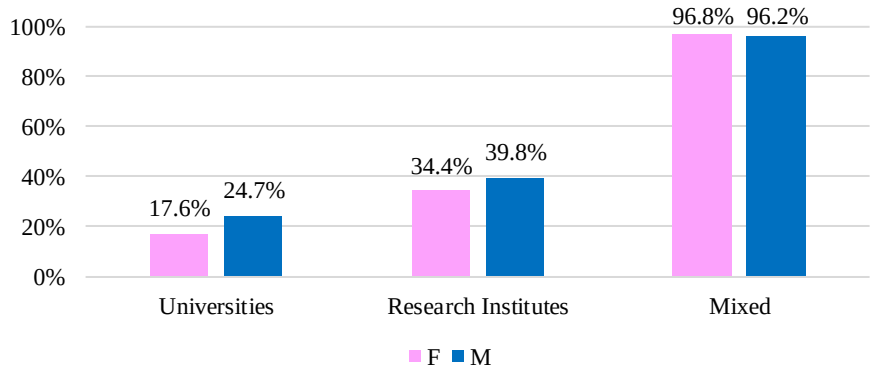


Figure 3. Share of Armenian professors with at least one 2017-2021 WoS publication, by gender and affiliation type.

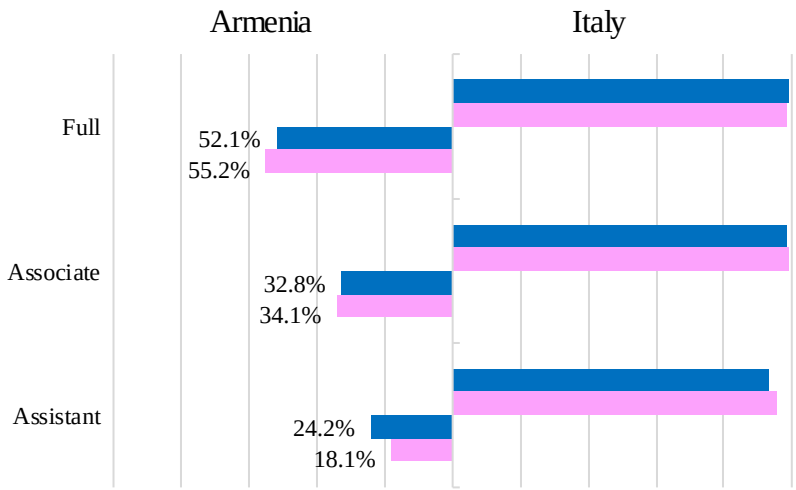


Figure 4. Share of professors with at least one 2017-2021 WoS publication, by gender, academic rank, and country.

Tables 4 and 5 provide insights into the yearly average output of professors (as measured by [2]) in both countries. Table 4 examines the entire dataset, while Table 5 focuses specifically on professors with at least one WoS publication. In Table 4, Italian professors exhibit significantly higher average yearly outputs compared to their Armenian counterparts across all fields. This difference is particularly pronounced in Clinical Medicine and Engineering, where Italian professors produce more than double the output of Armenian professors. Gender differences are also apparent, with male professors generally outperforming female professors in both countries. The only exceptions occur in Earth and Space Sciences (Armenia) and Physics (Italy).

Table 5 narrows the focus to active researchers, revealing gender disparities among those with at least one publication similar to those of the entire dataset. However, Armenia's gap between genders is lower than Italy's, in all fields but Mathematics and Physics.

Table 4. Yearly average 2017-2021 output of professors in the dataset, by gender, country, and field.

	Armenia			Italy		
Field	F	M	Δ(%)	F	M	Δ(%)
Biology	0.124	0.186		3.767	4.739	
Biomedical Research	0.159	0.202		5.530	8.502	
Chemistry	0.336	0.623		4.920	5.517	
Clinical Medicine	0.162	0.192		5.529	7.764	
Earth and Space Sciences	0.200	0.183		3.139	3.781	
Engineering	0.084	0.113		4.226	5.200	
Mathematics	0.091	0.325		1.746	2.208	
Physics	0.379	0.864		12.047	11.632	

Table 5. Yearly average 2017-2021 output of professors with at least one 2017-2021 WoS publication, by gender, country, and field.

	Armenia			Italy		
Field	F	M	Δ(%)	F	M	Δ(%)
Biology	0.560	0.829		3.804	4.790	
Biomedical Research	0.584	0.772		5.570	8.595	
Chemistry	1.018	1.417		4.949	5.562	
Clinical Medicine	0.592	0.619		5.639	7.926	
Earth and Space Sciences	1.021	0.717		3.220	3.881	
Engineering	0.557	0.611		4.313	5.299	
Mathematics	0.583	0.906		1.872	2.305	
Physics	0.968	1.702		12.588	11.955	

Fractional output

Tables 6 and 7 refine the analysis by focusing on fractional output (as measured by [3]), which adjusts for multi-authorship. With the exceptions of Clinical Medicine and Earth and Space Sciences, Armenian women exhibit lower fractional output compared to their male counterparts, even in fields with higher female participation, such as Biology and Biomedical Research. This suggests that while Armenian women are numerically well-represented in certain fields, their roles in collaborative projects may be less prominent, potentially limiting their overall fractional output. In contrast, in Italy, women's fractional output is always lower than men's across all fields, and differences between the two sexes are greater in Italy than in Armenia.

Table 6. Yearly average 2017-2021 fractional output of professors in the dataset, by gender, country and field.

	Armenia			Italy		
Field	F	M	Δ(%)	F	M	Δ(%)
Biology	0.023	0.037		0.618	0.803	
Biomedical Research	0.025	0.035		0.761	1.185	
Chemistry	0.080	0.143		0.825	0.996	
Clinical Medicine	0.031	0.029		0.847	1.204	
Earth and Space Sciences	0.046	0.033		0.638	0.781	
Engineering	0.020	0.034		0.963	1.209	
Mathematics	0.046	0.199		0.656	0.897	
Physics	0.103	0.205		0.787	0.976	

Table 7. Yearly average 2017-2021 fractional output of professors with at least one 2017-2021 WoS publication, by gender, country and field.

	Armenia			Italy		
Field	F	M	Δ(%)	F	M	Δ(%)
Biology	0.105	0.166		0.624	0.811	
Biomedical Research	0.092	0.136		0.767	1.198	
Chemistry	0.243	0.325		0.830	1.004	
Clinical Medicine	0.113	0.092		0.864	1.229	
Earth and Space Sciences	0.237	0.131		0.654	0.802	
Engineering	0.129	0.185		0.983	1.232	
Mathematics	0.297	0.555		0.703	0.937	
Physics	0.262	0.403		0.822	1.003	

Average impact

Tables 8 and 9 examine the average impact of professors' publication portfolios, measured by citation rates (as in [4]) and journal impact factors (as in [5]). Italian researchers, regardless of gender, outperform their Armenian counterparts in both metrics, reflecting Italy's more established academic infrastructure and global integration. In both countries, gender differences in average impact are field-dependent. For example, Armenian women in Clinical Medicine and Physics achieve higher average citation impacts than men, whereas their peers in Biomedical research and Biology show significantly lower averages. In contrast, Italian women in Biology, Chemistry, and Physics overcome men. This pattern underscores the interplay between field-specific norms and the visibility of women's research.

Table 8. Average impact of professors' publication portfolio, by gender, country, and field.

	Armenia			Italy		
Field	F	M	Δ(%)	F	M	Δ(%)
Biology	0.473	0.580		1.082	1.069	
Biomedical Research	0.326	0.651		1.084	1.093	
Chemistry	0.254	0.288		0.974	0.968	
Clinical Medicine	0.623	0.539		1.014	1.017	
Earth and Space Sciences	0.688	0.708		1.024	1.050	
Engineering	0.250	0.252		0.989	1.032	
Mathematics	0.286	0.292		0.883	0.951	
Physics	0.299	0.228		1.369	1.356	

Table 9. Average journal impact factor of professors' publication portfolio, by gender, country and field.

	Armenia			Italy		
Field	F	M	Δ(%)	F	M	Δ(%)
Biology	0.601	0.665		1.162	1.224	
Biomedical Research	0.474	0.557		1.137	1.143	
Chemistry	0.453	0.498		1.196	1.219	
Clinical Medicine	0.696	0.694		1.065	1.046	
Earth and Space Sciences	0.654	0.736		1.043	1.076	
Engineering	0.353	0.391		0.720	0.689	
Mathematics	0.405	0.386		0.947	0.944	
Physics	0.466	0.399		1.029	1.032	

Productivity

Figures 5 and 6 illustrate the distribution of research productivity (as measured by [1], transformed in percentiles), highlighting the disparities between Armenian and Italian professors and between genders within each country. Italian professors occupy higher productivity percentiles overall, with minimal gender differences. In Armenia, the distribution skews sharply, with a substantial proportion of women falling into lower productivity percentiles. However, among Armenian "productive" professors (those with at least one publication), the gender gap narrows slightly, suggesting that once structural barriers to productivity are overcome, women can achieve performance levels closer to those of their male counterparts.

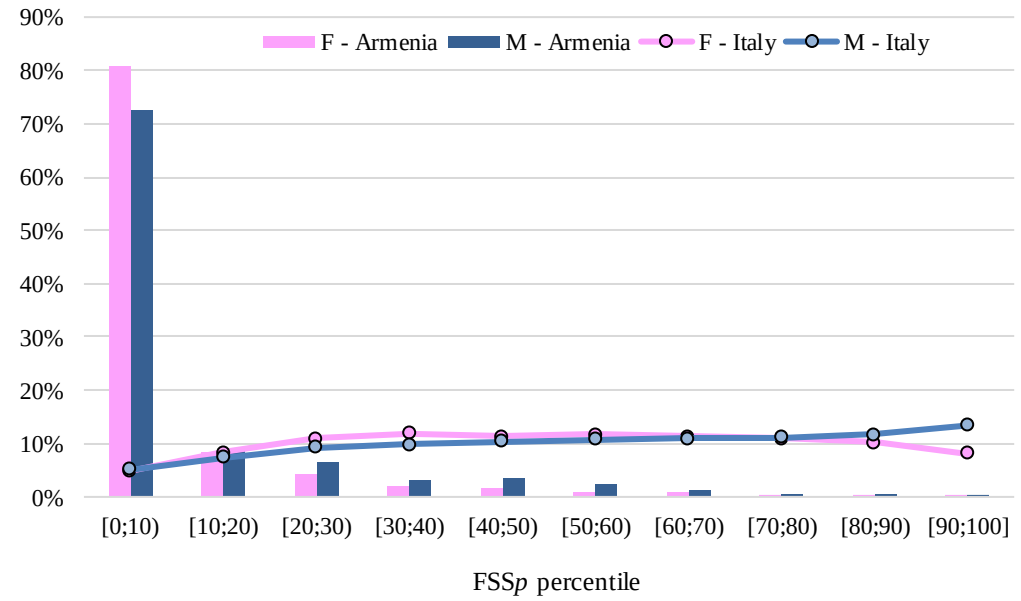


Figure 5. Distribution of 2017-2021 research productivity percentiles of Armenian and Italian professors, by gender.

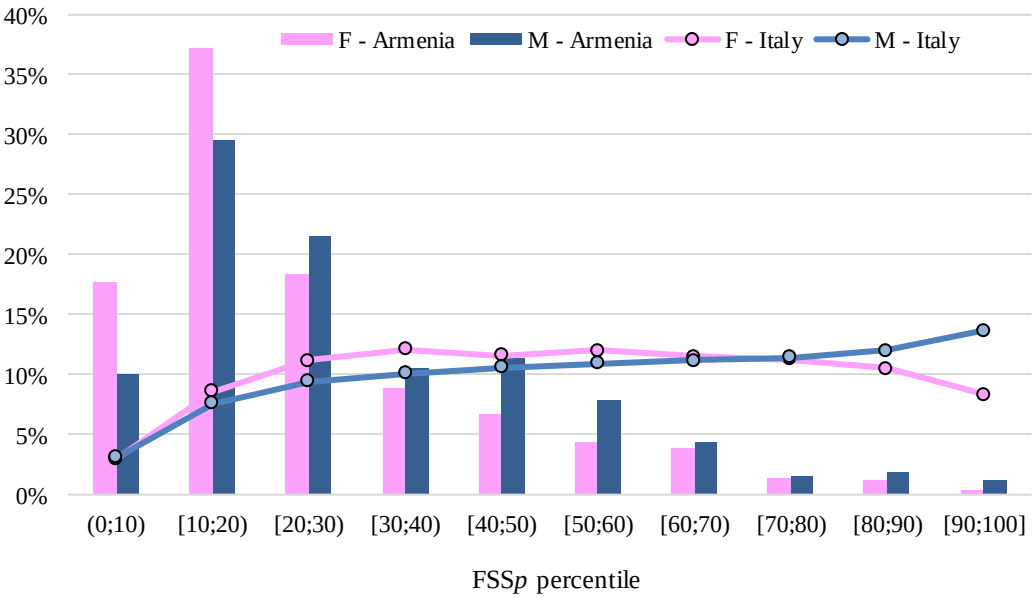


Figure 6. Distribution of 2017-2021 research productivity percentiles of Armenian and Italian "productive" professors, by gender.

Conclusions

This study has illuminated the complexities of gender disparities in academic research performance, using Armenia and Italy as case studies to explore how socio-cultural and institutional contexts influence the experiences of male and female researchers. The findings highlight not only the persistent gender gaps in both countries but also the ways these gaps differ due to structural and cultural factors.

Armenia presents a paradox: while it boasts a higher numerical representation of women in research (52 percent compared to Italy's 35.8 percent), this inclusivity does not translate into proportional research output or impact. Only 28.3 percent of Armenian researchers (with women consistently underrepresented in productive roles) have at least one WoS publication. In contrast, nearly all Italian professors (98.1%) are active in producing WoS-indexed publications, demonstrating a well-established academic system despite significant gender imbalances. These findings resonate with broader studies highlighting how numerical representation does not guarantee equity in access to resources or opportunities for advancement (Ceci & Williams, 2011; UNESCO, 2019).

In Italy, the research landscape demonstrates gendered hierarchies deeply embedded in academic structures. Women remain underrepresented in senior positions and produce fewer high-impact publications, consistent with global evidence showing that systemic biases, slower career progress, and disproportionate caregiving responsibilities hinder women's academic performance (Guarino & Borden, 2017; Abramo, D'Angelo, & Caprasecca, 2009). Nonetheless, Italian researchers benefit from robust academic networks and funding systems, which support higher productivity levels across genders compared to their Armenian counterparts.

The Armenian case, by contrast, underscores the challenges of a nascent research infrastructure compounded by traditional gender norms and systemic limitations, such as insufficient international collaboration and limited access to high-impact journals. Women, while numerically more represented in STEMM fields like Biology and Biomedical Research, face barriers in leadership roles and prominent collaborative opportunities. This finding aligns with studies from similar transitional contexts, where gender disparities are exacerbated by resource constraints and societal expectations (van den Besselaar & Sandström, 2016; Yeritsyan, 2019).

The results have significant implications for policy at both national and institutional levels.

In Armenia, interventions should prioritize enhancing research infrastructure and providing targeted support for women, such as mentorship programs, funding grants, and international exchange opportunities. Building capacity for international collaborations can mitigate structural barriers and increase the visibility of Armenian women researchers. This approach has proven effective in similar contexts, such as in Eastern Europe, where efforts to integrate into global research networks have reduced gender gaps (UNESCO, 2019).

In Italy, addressing the leaky pipeline in academic careers requires measures to ensure transparency in hiring, promotion, and funding allocation processes. Initiatives fostering work-life balance, such as flexible tenure-track models and family-friendly policies, could alleviate the career interruptions that

disproportionately affect women, as suggested by studies in other high-income countries (Borrell-Damián & Rahier, 2019; Stoet & Geary, 2018).

Both countries would benefit from fostering cross-disciplinary and international collaborations, particularly for women in male-dominated fields like Engineering and Physics, where barriers to entry and advancement are most pronounced. Research has shown that enhanced networking opportunities and visibility can significantly close productivity and impact gaps (Thelwall & Wilson, 2014; Caplar, Tacchella, & Birrer, 2017).

While this study provides valuable insights, several limitations must be acknowledged. The bibliometric analysis relies on WoS-indexed publications, potentially underestimating contributions in non-indexed or local-language journals, particularly in Armenia. Field-specific norms, such as collaborative practices and citation behaviors, may also influence the observed gender disparities and require further investigation. Moreover, cross-country comparisons are complicated by structural differences in academic systems—e.g., the broader inclusion of Armenian research staff versus the exclusive focus on university professors in Italy.

Future research should integrate qualitative methods to capture the nuanced interplay of cultural, institutional, and individual factors shaping gender disparities in academia. Comparative studies involving additional countries and disciplines could further elucidate how national policies and practices foster or hinder gender equity in academic research.

While numerical representation is an important starting point, achieving true gender equity in research requires systemic changes to address entrenched biases and structural barriers. The findings from this study contribute to a growing body of evidence advocating for targeted, context-sensitive interventions to create inclusive academic environments.

References

Abramo, G., & D'Angelo, C.A. (2014). How do you define and measure research productivity? *Scientometrics, 101*(2), 1129–1144.

Abramo, G., Aksnes, D.W., & D'Angelo, C.A. (2020). Comparison of research productivity of Italian and Norwegian professors and universities. *Journal of Informetrics, 14*(2), 101023.

Abramo, G., Aksnes, D.W., & D'Angelo, C.A. (2021). Gender differences in research performance within and between countries: Italy vs Norway. *Journal of Informetrics, 15*(2), 101144.

Abramo, G., D'Angelo, C.A., & Caprasecca, A. (2009a). Gender differences in research productivity: A bibliometric analysis of the Italian academic system. *Scientometrics, 79*(3), 517–539.

Abramo, G., D'Angelo, C.A., & Caprasecca, A. (2009b). The contribution of star scientists to overall sex differences in research productivity. *Scientometrics, 81*(1), 137–156.

Abramo, G., D'Angelo, C.A., & Di Costa, F. (2019). A gender analysis of top scientists' collaboration behavior: Evidence from Italy. *Scientometrics, 120*(2), 405–418.

Abramo, G., D'Angelo, C.A., & Felici, G. (2019). Predicting long-term publication impact through a combination of early citations and journal impact factor. *Journal of Informetrics, 13*(1), 32–49. https://doi.org/10.1016/j.joi.2018.11.003

Abramo, G., D'Angelo, C.A., & Mele, I. (2022). Impact of Covid-19 on research output by gender across countries. *Scientometrics, 127*(12), 6811–6826.

Abramo, G., D'Angelo, C.A., & Murgia, G. (2013). Gender differences in research collaboration. *Journal of Informetrics, 7*(4), 811–822.

Abramo, G., D'Angelo, C.A., & Rosati, F. (2016). Gender bias in academic recruitment. *Scientometrics, 106*(1), 119–141.

Addis, E., & Villa, P. (2003). The editorial boards of Italian economics journals: A gender analysis. *Feminist Economics, 9*(1), 75–84.

Aksnes, D.W., Rorstad, K., & Sivertsen, G. (2011). Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of the American Society for Information Science and Technology, 62*(4), 628–636.

Aksnes, D.W., & Sivertsen, G. (2019). A criteria-based assessment of the coverage of Scopus and Web of Science. *Journal of Data and Information Science, 4*(1), 1–21.

Alper, J., & Gibbons, A. (2017). The leaky pipeline in academia: Barriers to gender equality. *National Academies Press.*

Bendels, M.H.K., Müller, R., Brueggmann, D., & Groneberg, D.A. (2018). Gender disparities in high-quality research revealed by Nature Index journals. *PLOS ONE, 13*(1), e0189136.

Bettio, F., & Verashchagina, A. (2009). Gender segregation in the labour market: Root causes, implications, and policy responses in the EU. *European Commission's Expert Group on Gender Equality and Employment.*

Borrell-Damián, L., & Rahier, M. (2019). Women in university leadership: Subtle leaks in the pipeline to the top. *European University Association.* https://www.eua.eu/our-work/expert-voices/women-in-university-leadership-subtle-leaks-in-the-pipeline-to-the-top.html

Bozeman, B., & Corley, E. (2004). Scientists' collaboration strategies: Implications for scientific and technical human capital. *Research Policy, 33*(4), 599–616.

Caplar, N., Tacchella, S., & Birrer, S. (2017). Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy, 1,* 0141.

Ceci, S.J., & Williams, W.M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences, 108*(8), 3157–3162.

Dworkin, J.D., Linn, K.A., Teich, E.G., Zurn, P., Shinohara, R.T., & Bassett, D.S. (2020). The extent and drivers of gender imbalance in neuroscience reference lists. *Nature Neuroscience, 23*(8), 918–926.

Elsevier. (2020). *The research gender gap: Research performance by gender across 15 countries.* Elsevier Gender Report.

Guarino, C.M., & Borden, V.M.H. (2017). Faculty service loads and gender: Are women taking care of the academic family? *Research in Higher Education, 58*(6), 672–694.

Huang, J., Gates, A.J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences, 117*(9), 4609–4616.

Larivière, V., Archambault, É., Gingras, Y., & Vignola-Gagné, É. (2006). The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology, 57*(8), 997–1004.

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C.R. (2013). Bibliometrics: Global gender disparities in science. *Nature, 504*(7479), 211–213.

Lena, G. (2023). Women in academia: Why and where does the pipeline leak, and how can we fix it? *MIT Science Policy Review, 4*, 102–109.

Mairesse, J., & Pezzoni, M. (2015). Does gender affect scientific productivity? Evidence from a large-scale French survey. *Revue économique, 66*(1), 65–113.

Moscatelli, M., et al. (2019). Academic career progressions and gender gaps: Italy in a European context. *Gender in Management, 34*(3), 233–249.

Stoet, G., & Geary, D.C. (2018). The gender-equality paradox in science, technology, engineering, and mathematics education. *Psychological Science, 29*(4), 581–593.

Thelwall, M., & Wilson, P. (2014). Gender differences in bibliometric indicators and implications for career development. *Journal of Informetrics, 8*(2), 292–305.

Tsakanova, G., Arakelova, E., Matevosyan, L., Petrosyan, M., Gasparyan, S., Harutyunyan, K., & Babayan, N. (2021). The role of women scientists in the development of ultrashort pulsed laser technology-based biomedical research in Armenia. *International Journal of Radiation Biology, 98*(3), 489–495. https://doi.org/10.1080/09553002.2021.1987566

UNDP Armenia. (2020). *Gender equality brief: Progress and challenges*. United Nations Development Programme Armenia.

UNESCO (2019). *Global Education Monitoring Report 2019: Gender report*. UNESCO Publishing.

UNESCO Institute for Statistics. (2020). *Fact sheet no. 60: Women in science*. Retrieved from http://uis.unesco.org/sites/default/files/documents/fs60-women-in-science-2020-en.pdf

van den Besselaar, P., & Sandström, U. (2016). Gender differences in research performance and its impact on careers: A longitudinal case study. *Scientometrics, 106*(1), 143–162.

Witteman, H.O., Hendricks, M., Straus, S., & Tannenbaum, C. (2019). Are gender gaps due to evaluations of the applicant or the science? A natural experiment at a national funding agency. *The Lancet, 393*(10171), 531–540.

Yeritsyan, S. (2019). Gender issues in higher education: Armenia's progress and ongoing challenges. *Armenian Journal of Social Sciences, 5*(2), 97–112.

Zazyan, M. (2009). Armenian women in physics. *AIP Conference Proceedings, 1119*, 77–78. https://doi.org/10.1063/1.3137915

Zazyan, M. (2019). Women physicists in Armenia: Why so few? In G. Cochrane, C. Singh, & N. Wilkin (Eds.), *Women in Physics: 6th IUPAP International Conference on Women in Physics, AIP Conference Proceedings 2109* (pp. 0500). American Institute of Physics.

https://doi.org/10.51408/issi2025_021

Gendered Collaboration Networks and Their Consequences on Conflicts between Academics

László Lőrincz[1], Brigitta Németh[2], Tamás Felföldi[3]

[1]*laszlo.lorincz@uni-corvinus.hu*
Institute of Data Analytics and Information Systems & ANETI Lab, Corvinus Institute for Advanced Studies, Corvinus University of Budapest, H-1093 Budapest, Fővám tér 8 (Hungary)
ANETI Lab, Institute for Economics, HUN-REN Centre for Economic and Regional Studies, H-1097 Budapest, Tóth Kálmán utca 4 (Hungary)

[2]*nemeth.brigitta@krtk.hu*
Doctoral School of Economics, Business and Informatics, Corvinus University of Budapest, H-1093 Budapest, Fővám tér 8 (Hungary)
ANETI Lab, Institute for Economics, HUN-REN Centre for Economic and Regional Studies, H-1097 Budapest, Tóth Kálmán utca 4 (Hungary)

[3]*tamas.felfoldi@proton.me*
Corvinus University of Budapest, H-1093 Budapest, Fővám tér 8 (Hungary)

Abstract

In academia, a major field of knowledge production, the quality of interactions among coworkers plays a critical role. Negative ties, such as conflicts or avoidance between researchers, can impede the progress of research projects, and adversely affect individual career advancement. Our study reveals that in the academic sector, these negative relations are not "gender-neutral" as they are experienced 47% more often by women. Using a large representative survey linked to official scientometric records from Hungary, we demonstrate that the conditions under which women and men experience these negative relations differ. Women experience more conflicts if they act as brokers in scientific collaborations, and fewer conflicts when they are members of cohesive groups. These factors, however, do not influence the number of conflicts for men. Thus, we can argue, that while being in a broker position holds the promise for scientific success, it comes with the price of more workplace conflicts for women. Regarding the role of gender diversity and cross-gender collaboration, we find that women report less negative relations when they collaborate with fellow women if they are a small minority, but in diverse fields, cross-gender collaboration comes with fewer conflicts.

Introduction

Employees are involved in negative relationships in workplaces, such as animosity, avoidance, or exclusion. While positive relations, such as support increase satisfaction with workplace relationships, negative ones decrease it, indirectly decreasing the attachment to the workplace too (Venkataramani et al., 2013). Similar outcomes are reported concerning performance. Employees who are more central in advice networks tend to be more efficient, while centrality in hindrance networks is inversely related to performance. The extent of hindrance relationships also harms performance on the group level (Sparrowe et al., 2001). Moreover, there is evidence that negative relations are more influential than positive ones in one's network (Kane & Labianca, 2006).

In academia, as a major field of knowledge production, the quality of interactions among coworkers plays a critical role, from collaboration to talent recruitment. Negative ties, such as conflicts or avoidance between coworkers, can hinder the progress of research projects, and may negatively influence individual career progression.

What is more interesting, however, is that in academia, negative relations do not seem to be "gender-neutral", that is they are unproportionally more often experienced by women. In the qualitative study of (Gersick et al., 2000) women reported negative aspects of their relationships more than four times higher than men. In light that in non-academic contexts (business organizations) the number of negative ties was found non-different between genders (Merluzzi, 2017), if such an excess number of difficult relationships are experienced by women in academia, it can be a significant liability for female researchers, a factor contributing to the observed higher dropout rate of female scientist (Lietz et al., 2024).

If such gender differences exist, it is also interesting that under what conditions women (and men) experience them. A potential argument is that the gender differences in negative tie formation are influenced by gender norms that convey gender-specific behaviors and expectations in organizations (Ridgeway, 2009, 2001; Elsesser & Lever, 2011; Eagly & Karau, 2002; Heilman & Okimoto, 2007). Furthermore, the effect of proportions, or the relative number of socially, culturally, or biologically different people in a group, has been also observed to be significant in shaping these role expectations (Benan & Olca, 2020; Holgersson & Romani, 2020; Kanter, 1977; Zimmer, 1988). Thus, highlighting the amplifying effect of (low) diversity in a work environment is crucial to understand the gender differences in conflict-type relationships.

Given the intertwined nature of diversity and role expectations, we aim to explore the effect of field diversity on the relationship between conflicts, social capital, and the role of male or female weight in the collaboration network. Our goal is to enrich the gender-focused literature on negative ties in the professional networks of the scientific workforce by exploring the association between gender role incongruent social capital and conflicts in balanced and male-dominated fields.

For the analysis, we use survey data linked to administrative scientometric data from Hungary. In this country, the gender ratio of scholars in different disciplines varies between a very low share of females, 12.6% in Engineering, to almost perfectly balanced (52,4%) in Literature and Linguistics, which makes the Hungarian scientific sector a good setting to explore the consequences of low diversity and tokenism regarding conflict type relationships.

Our results reveal that there is indeed a significant difference between male and female academics in the number of negative relations; women report 47% more negative relations, and the difference remains significant after controlling for individual attributes. We also find that the factors predicting the number of negative relations are different for men and women. Women experience more conflicts if they act as brokers in scientific collaborations, and less if they are members of cohesive groups. These factors however do not influence the number of conflicts for men. Thus, we can argue, that while being in a broker position has the promise of scientific

success (Guan et al., 2017; Jadidi et al., 2018), this comes with the price of more workplace conflicts for women. About the moderating role of gender diversity in the field, we find that in comparison to male-dominated fields, in more balanced fields men report more conflicts and women report less. Interestingly, however, women experience more conflicts if they collaborate with other women in balanced fields.

Theory and hypotheses

In every organization, but in academia particularly, success depends on collaboration. Positions in the networks of social relations represent specific advantages and at the same time, different characteristics are attributed to the holders by the others. Social capital as cohesion, for instance, is about strong ties: close relationships characterized by trust, cooperation, mutual support, or solidarity (Coleman, 1988, 1990) and means that everyone knows each other in one's network. In academia, these strong ties can manifest as long-lasting research collaborations (Dahlander & McFarland, 2013), and more cohesive networks of scientists are found to be more productive, but only in already well-established fields (Jansen et al., 2010). Social capital as brokerage is captured in networks with sparsely connected parties by brokers, who bridge these gaps (structural holes), which allow movement in versatile social circles with access to non-redundant information through these weak ties (Burt, 1992; Barthauer et al., 2016). Therefore, weak ties were shown to contribute to success in different fields (Fronczak et al., 2022; Rajkumar et al., 2022), and it has also shown that access to more unique information is a key mechanism beyond this success in organizations (Aral & Dhillon, 2023; Gonzalez-Brambila, 2014). In academia, being in a broker position also tends to have a positive influence on scientific success (Guan et al., 2017), especially for junior scholars (Patel et al., 2019). Moreover, the positive influence of brokerage on academic success was not found to be moderated by gender (Jadidi et al., 2018).

Interaction dynamics in organizations is also driven by gender norms (Ridgeway, 2001, 2009; Elsesser & Lever, 2011; Eagly & Karau, 2002; Heilman & Okimoto, 2007). Ridgeway (2009) argues that gender is a primary cultural frame for coordinating behavior and organizing social relations; thus, it shapes organizational structures as well. The stereotypical female gender roles are communal roles, including nurturing, caring, and sensitivity. Male roles are more agentic, like ambitious, assertive, and direct (Elsesser & Lever, 2011). Eagly and Karau (2002) propose that acting incongruent with these stereotypical roles leads to being evaluated negatively. Heilman and Okimoto (2007) confirm that women indeed face penalties for success in traditionally male domains if they lack nurturing and socially sensitive communal attributes. At the same time, the control benefits and relative independence associated with a broker's position are congruent with gender role expectations for men, they are not congruent with gender role expectations for women (Eagly, 1987).

About gender differences in navigating organizational networks, Burt (1998) finds that women benefit from different network strategies than their male coworkers to achieve success. While successful men are more likely to be brokers in networks with more structural holes, successful women are involved in networks characterized

by few structural holes and higher cohesion since they have less legitimacy in a work setting. Contrary to this, Ibarra (1997) finds that high-performing women are more likely to have connections outside their organizational units. The results of Lutter (2015) analyzing the film industry align with this, showing that women suffer a career penalty if they work in cohesive groups. Carboni and Gilman (2012) however add that women are more likely to experience social stress when they occupy brokerage positions and so attempt to address often conflicting expectations of the relationship partners from disconnected social groups. Jadidi et al. (2018) confirm that women are more likely to embed into networks with higher cohesion and to have lower brokerage than men in academia as well. An opposite tendency, that women have more brokerage was reported by Barthauer et al. (2016). Note that they consider mentorship networks, while Jadidi et al. (2018) considered scientific collaboration networks.

Based on the arguments that communal roles are typically associated with females and agentic roles with males, and the observation that women tend to occupy more cohesive and less bridging positions in organization networks, we expect that:

H1: Being in bridging positions in the scientific collaboration network will be positively associated with conflict relationships for women, however, we expect no such association for men, reflecting gender role expectations.

Structural constraints however interact with role expectations, and they were found to jointly determine these gender specificities regarding social capital (Eagly & Karau, 2002; Rudman & Glick, 2001; Carboni, 2023). The concept of "tokenism" has been widely used to explain women's experiences, such as role entrapment, as they enter traditionally male occupations and represent a clear minority in an organization. 15% is the estimated critical ratio of the minority group to apply the category (Benan & Olca, 2020; Holgersson & Romani, 2020; Kanter, 1977; Zimmer, 1988). Schoen et al. (2018) emphasize the altering effect of diversity in this context; when women are in token situations, they benefit from networks with few structural holes, and their male colleagues benefit from networks with many structural holes. While in non-token situations, i.e., when the proportion of women exceeds 15%, men and women benefit from the same network structures.

We therefore expect that being in a token situation contributes to lower acceptance by peers if not following the gender-specific role expectations thus engaging in bridging collaborations.

H2: We propose that women with high brokerage experience fewer conflicts in diverse fields.

Concerning gender in organizations, homophily is also an important driver of network relations. McPherson et al. (2001) define it as "the principle that contact between similar people occurs at a higher rate than among dissimilar people". In academia, homophily in the collaboration patterns was analyzed by Kwiek and Roszka (2021). They found gender homophily to apply to male scientists—but not to females. The majority of male scientists collaborate solely with males, while all-female collaboration was found to be marginal. Yap and Harrigan (2015) highlight the significance of homophily in understanding negative tie formation, noting that

men and women tend to direct negative ties towards the opposite gender rather than their own.

At the same time, the effect of proportions, or the relative number of socially and culturally different people in a group, has been observed to be significant in shaping social networks as well (Kanter, 1977). Experiments by Szell and Thurner (2013) have shown that in a male-dominated context, men discount women as legitimate competitors, which would reduce the odds of work conflict with women initiated by men. They observe male competitiveness exclusively among themselves, with fewer cooperative links between them and a reluctance to reciprocate hostile actions from females. On the contrary, females exhibit stronger homophily and network closure among themselves in their collaboration networks in the analyzed online gaming environment. Ely (1994) finds that the principal mechanism through which the representation of women influences their relationships is social identity. Women were less likely to experience gender as a positive basis for identification in organizations with few senior women and less likely to perceive senior women as role models with legitimate authority. When being in a small minority, women are more likely to perceive competition in female peers instead of finding support in these relationships (Duguid, 2011), and apply masculine self-descriptions themselves (Derks et al., 2011). Merluzzi (2017) adds that even though men and women were equally likely to cite a negative work relationship, women were more inclined than men to cite a negative relationship with another woman if they had no female social support in the workplace network. Being a minority in categories such as race, gender, and age can generate conflict according to Pelled (1996) as well. Being different from other group members may negatively shape a person's perspective on group interactions but it is also possible that having a demographically distinct group member truly fosters conflict.

These mechanisms lead to different hypotheses regarding the relationship between the number of women in the field and the number of conflicts by gender. For men, the increased number of women may contribute to more conflicts based on homophily theory; while men do not see women as competitors in token situations, if the female ratio increases, this would change.

H3: We expect the number of conflicts to be positively related to the ratio of the other gender in the scientific field for men.

On the other hand, for women:

H4: The number of conflicts will be negatively related to their ratio in the scientific field for women, indicating the influence of positive social identification and so the lower level of competitiveness.

Besides, we are investigating whether there is an interaction between the gender ratio of individuals' collaboration networks and the gender ratio of the scientific field. We consider that while the gender makeup of the field is something external, academics in minority positions might opt to collaborate with others in similar situations to mitigate their position and seek support. This implies that collaborating with co-authors isn't just a strategic career move, but also a way to find support through connections with people of the same gender.

H5: If women collaborate more with other women in male-dominated fields, (or men collaborate with fellow men in female-dominated fields) they experience fewer conflicts.

Data

For our analysis, we use survey data linked to administrative scientometric data on the individual level. The survey was initiated by the Hungarian Young Academy about working conditions, income, satisfaction, international mobility, and professional relations of Hungarian researchers under the age of 45. Its special feature was that respondents were asked if they consented to link their scientometric data in the Hungarian scientometric system (MTMT) to the survey and to provide their IDs for the linking.

The survey was conducted online. Invitations were sent by the Hungarian Academy of Sciences to all members of the Academy's public body under the age of 45, and to all researchers who have defended their Ph.D. after 1992 and gave active consent to receive science-related news. In addition, a Facebook campaign supported the recruitment of respondents. The data was collected in September–October 2021.

The number of completed responses was 1,219, of which we were able to analyze the responses of 1,135 respondents after data cleaning. Linking the data was possible for 1,009 individuals. The linked database was deposited and only accessible at the Data Bank of the Centre for Economic and Regional Studies, providing a secure on-site environment for analysis.

Measures

For the measurement of conflicts, we used the following name-generator question from the survey: "Do you have someone with whom you have a difficult or burdensome relationship, perhaps even conflict from time to time?" Respondents could provide a list of up to ten names or pseudonyms in the form. We consider the number of names provided as the number of conflict (or adversary) relationships. We use these terms as synonyms in the analysis.

As for the independent variables, we used self-reported survey data on gender, academic rank, and discipline. Academic rank was measured using four categories corresponding to the Hungarian standards: (1) Ph.D. Student or assistant lecturer (2) Assistant professor or research fellow (3) Associate professor or senior research fellow (4) Full professor. Discipline was measured according to the eleven-class classification of the Hungarian Academy (see Figure 2B).

We measured the gender ratio of the specific fields using the directories of the Academy's public body. Membership lists by disciplines were retrieved from its official website[1] in 2023, and genders were identified by matching with the list of registerable surnames in Hungary[2], and by using the 'gender' R package of Mullen (2021)[3] for non-Hungarian surnames. The share of female scientists varied

[1] https://mta.hu/koztestuleti_tagok, N=17,428
[2] https://nytud.hu/en/oldal/utonevjegyzek
[3] https://cran.r-project.org/web/packages/gender/readme/README.html

significantly by field, from 13% in Engineering to 52% in Language and Literature (Figure 1B).

Another group of measures was calculated from the scientometric data. Considering citations, we used citation values standardized by academic age (time since the first publication) and discipline.

To measure collaboration characteristics, we first created a weighted collaboration network between scientists. We defined weights between two scientists taking into account that co-authorship on a paper with many authors indicates a weaker collaboration than co-authorship on a paper with only two authors, following (Newman, 2001):

$$w_{ij} = \sum_k \frac{\delta_i^k \delta_j^k}{n_k - 1}$$

where δ_i^k is the indicator that person *i* is the author on paper *k* and n_k is the number of authors of paper *k*. Having the weighted collaboration network, we calculated the following measures. *(Strength of) Collaboration with women* is calculated as:

$$Collab\ w.women_i = \frac{\sum_j w_{ij \mid gender(j) = female}}{\sum_j w_{ij}}$$

Burt's constraint measure (Burt, 1992) measures the redundancy of ties and, thus is used as an inverse measure for bridging position in the network:

$$C_i = \sum\nolimits_{j \in V_{i, i \neq j}} \left(\sum\nolimits_{q \in V_{i, q \neq i, j}} p_{ij} + p_{iq} p_{qj} \right)^2$$

where p_{ij} are proportional tie strengths, defined as

$$p_{ij} = \frac{w_{ij} + w_{ji}}{\sum_{q \in V_{i, q \neq i}} (w_{iq} + w_{qi})}$$

and V_i is the ego-network of person *i*.

The weighted clustering coefficient. In general, local clustering measures the intensity of closed triangles around the individual that indicate structural embeddedness (Nahapiet & Ghoshal, 1998). We use its implementation applied for weighted networks, as suggested by (Onnela et al., 2005):

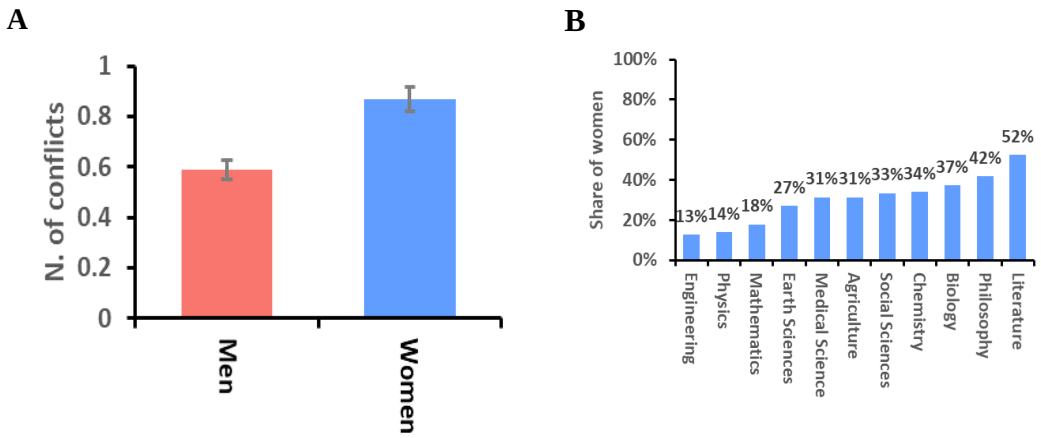
$$\tilde{C}_i = \frac{2}{k_i(k_i - 1)} \sum_{j,k} \sqrt[3]{\tilde{w}_{ij} \tilde{w}_{jk} \tilde{w}_{ki}}$$

where

$$\tilde{w}_{ij} = \frac{w_{ij}}{\max(w_{ij})}$$

Descriptive results

Figure 1A indicates that on average, women report significantly, about 40% more conflict relations. If we do not control for any other differences, men are more likely to be highly cited than women (Figure 1C), a result that is congruent with large scientometric studies (Huang et al., 2020; Larivière et al., 2013; Meho, 2022). In terms of scientific collaboration patterns, we see a substantive tendency of homophily, that women are more likely to collaborate with women than men do, therefore women have higher average strength of collaboration with women (Figure 1D). By visualizing it by scientific field (Figure 1E), we see that this is largely due to induced homophily because women collaborate with other women more often in fields where the ratio of female researchers is higher. However, we can observe choice homophily too, indicated by that women are more likely to collaborate with women than men are in every scientific field. Being involved in closed communities (measured by clustering), versus creating bridges in the collaboration networks (measured by constraint), is not different by gender (Figure 1F-G), similar to what Schoen et al. (2018) reports. Thus we neither observe that female researchers would have more brokerage capital than men do (in contrast to Barthauer et al., 2016), nor that they would have less (in contrast to Jadidi et al., 2018). In this aspect of the number of coauthors (degree), we do not observe gender differences either (Figure 3H), similar to Bozeman & Gaughan (2011) and Zeng et al. (2016).



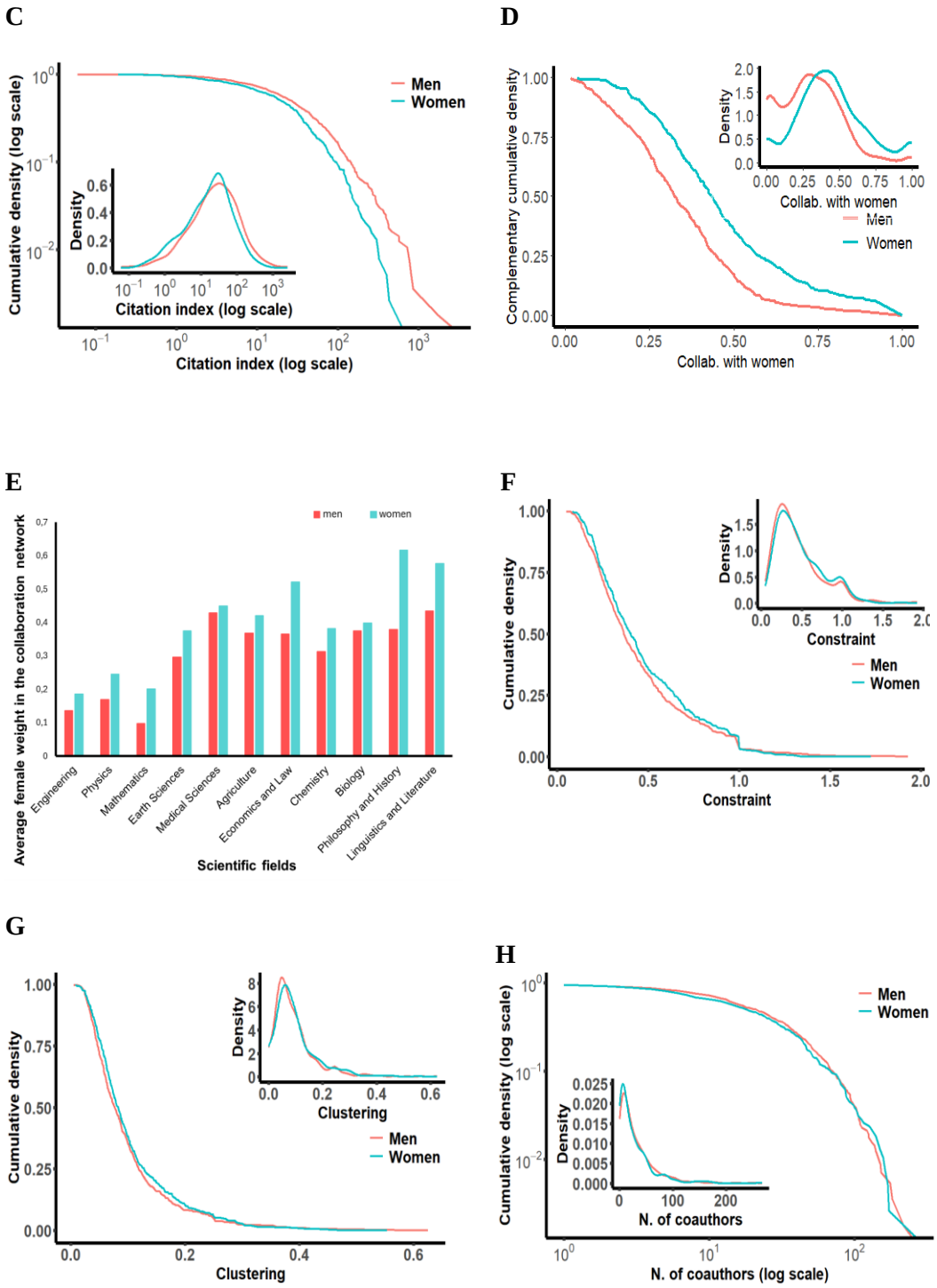


Figure 1. Descriptive statistics. A. Number of conflicts reported by gender (mean and SEM). B. Share of women in scientific fields. C. Normalized citation by gender. D. Collaboration strength with women by gender. E. Collaboration strength with women by academic field and gender. F. Burt's constraint by gender. G. Weighted clustering coefficient by gender. H. Number of coauthors by gender.

Results for statistical tests of hypotheses

We start the presentation with the statistical analysis of the relationship between ego-network position and the number of conflict relations. As the number of conflicts is a count variable, we model it using a Poisson regression. Our key independent variables corresponding to Hypothesis 1 are those describing whether ego-networks are structurally cohesive (that we measure by weighted clustering), or the individual is in a broker position (that we measure by the inverse of constraint). As these variables are highly correlated (Figure 2), we examine them in separate models. Important control variables in the regression are indicators of academic rank, citations, and the number of co-authors. The number of coauthors is important because clustering and constraint measures are empirically correlated with degree in most social networks (Marsden, 1990; Newman, 2003). Academic rank and citations are important cofounders, as they tend to be different by gender, and they are also correlated with the structural position of the researcher (more successful and senior researchers have more publications and more open network positions compared to beginners) (Figure 2).

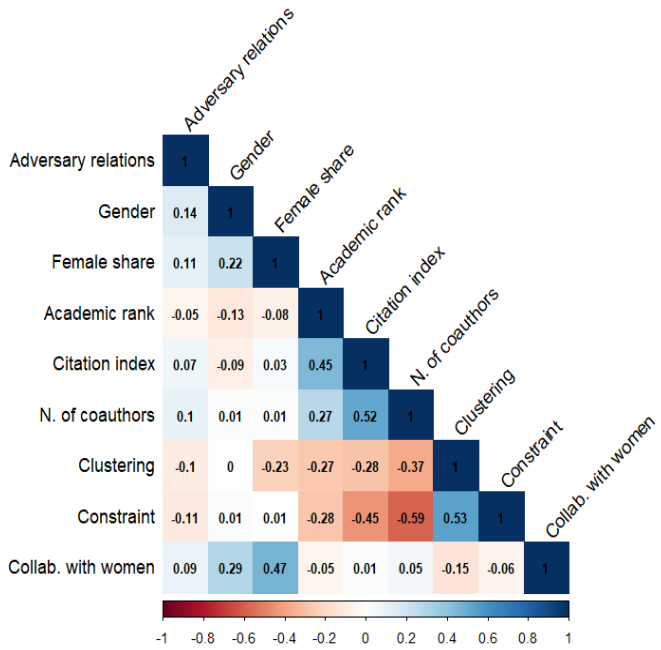


Figure 2. Correlation matrix of the variables used in our models.

Figure 3 displays the coefficients of Poisson regressions with the number of adversary relations as the dependent variable. When we consider men and women together (Figure 3A), we see that the most important predictor of the number of conflicts is gender; women have more conflicts. In addition, academic rank and clustering are also significant, indicating that more senior researchers and those who are more embedded in the coauthor network (higher clustering) have somewhat fewer conflicts. The coefficients of the scientific field dummies are not significant.

If we consider men and women separately (Figure 3B), we see that the factors predicting the number of conflicts are largely gender dependent. Men have more conflicts if they have more coauthors and if they are cited more, thus, we might say that their conflicts are related to their success in publications. These factors are in turn not significant for women. They have more conflicts if they are less embedded in co-author networks and if they have more junior ranks. These factors however are not significant for men. In Figure 3C we replace the clustering measure with Burt's constraint measure, which captures the inverse of bridging positions in networks. We see that the results are consistent with the previous panel. Women have more conflicts if their constraint is low (when they are in bridging positions), while the constraint is not significant for men. Taken together, we see that if women occupy bridging positions in co-author networks, connecting people who are otherwise unconnected, they experience more conflicts. However, if they are in closed networks, where everyone works with everyone else, it prevents them from conflicts. For men, however, clustering and constraint are not significant. This is what we expected in Hypothesis 1.

In Hypothesis 2 we put forward that being in a broker position creates more conflicts for women, especially if they are in token positions. Thus, to test this hypothesis, we replace the field dummies with the share of women in the corresponding scientific field and add its interaction term with the constraint measure. Because the data on the share of women is an aggregate by scientific fields, and the other variables are observed for the individuals, we use a multilevel (random intercept) specification of the Poisson regressions in this case. The coefficients of these models are displayed in Figure 3D. It is visible that although the coefficient of the "Female share x Constraint" interaction is positive, as expected in H2, it is not significant. Therefore, we could not justify that women in broker positions (having lower constraint) would have less conflict relationships in more balanced fields (if the share of women increases).

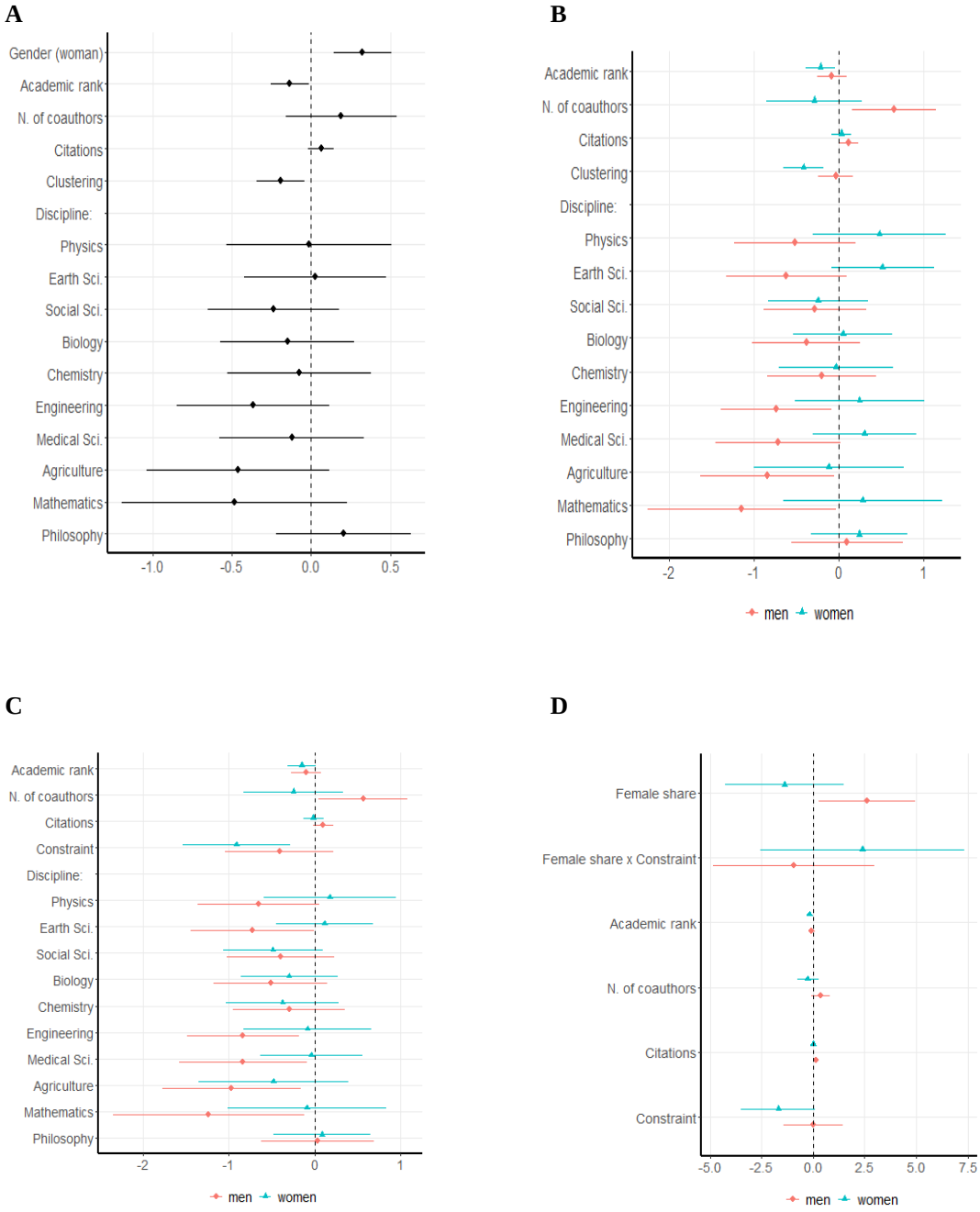
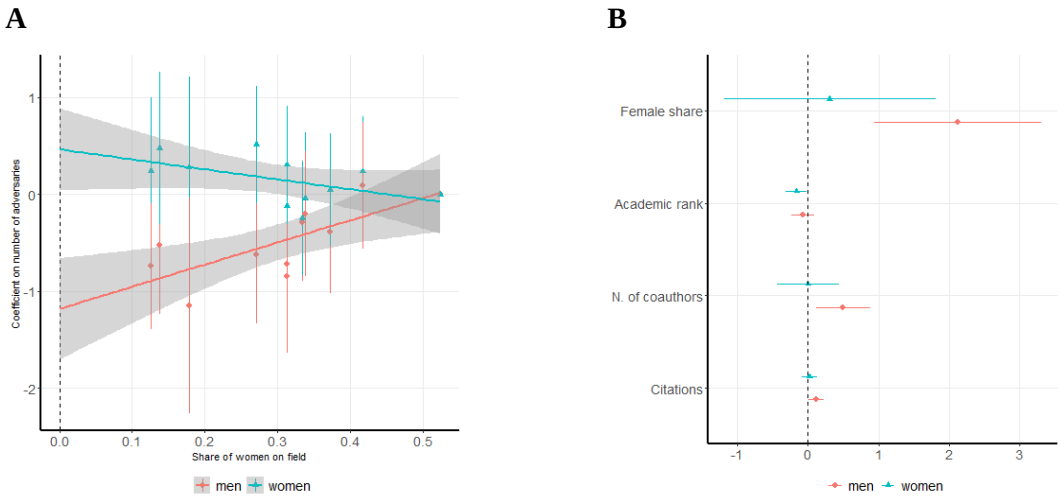


Figure 3. Results of Poisson regressions on the number of conflict relations (coefficients and confidence intervals). A. Both genders together B. Separate models by gender C. Separate models by genders, using Burt's constraint instead of clustering. D. Separate models by gender, using the share of women on the field and its interaction. Notes. N= 422 men + 301 women. Scale of variables: N. of coauthors (100), Clustering (0.1).

In Figure 4 we focus on the impact of gender composition on the number of conflicts. In Figure 4A we order the coefficients of scientific fields according to the share of women on the fields on the *x-axis* from the model presented in Figure 3C. We can observe a tendency that men have fewer conflicts in male-dominated fields, while they experience conflicts in fields with more balanced gender compositions. However, women tend to have more conflicts if they are in a token position, in contrast to more balanced fields. We test this tendency statistically in Figure 4B, where we replace the field dummies with the share of women in the corresponding scientific field. Because the data on the share of women is an aggregate by scientific fields, and the other variables are observed for the individuals, we use a multilevel (random intercept) specification. The coefficient plot shows that the share of women significantly increases the number of conflicts for men, corresponding to Hypothesis 3, but the tendency that women have fewer conflicts if they are not in token position (corresponding to Hypothesis 4) is not significant.
Figure 4C adds the interaction term of the collaboration with women and the share of women on the field, which is significant for women but not for men. The first conclusion is that men experience fewer conflicts, if they work in a male-dominated field, and it is not related to their collaboration patterns with men or women. Second, women experience more conflicts, if they collaborate with men in a male-dominated field, or if they collaborate with women in a more balanced field, which supports Hypothesis 5. From the point of view of the female scientists, our results suggest that both collaborating between females if they are a small minority or collaborating with men in gender-diverse fields can reduce conflicts. This tendency is visualized in Panel D in terms of the predicted change in the number of adversaries for the lowest and the highest observed female ratios as an example. For males, however, we cannot verify Hypothesis 5, as we only have balanced and male-dominated fields in the data, thus do not have ones, where men would be in minority position.



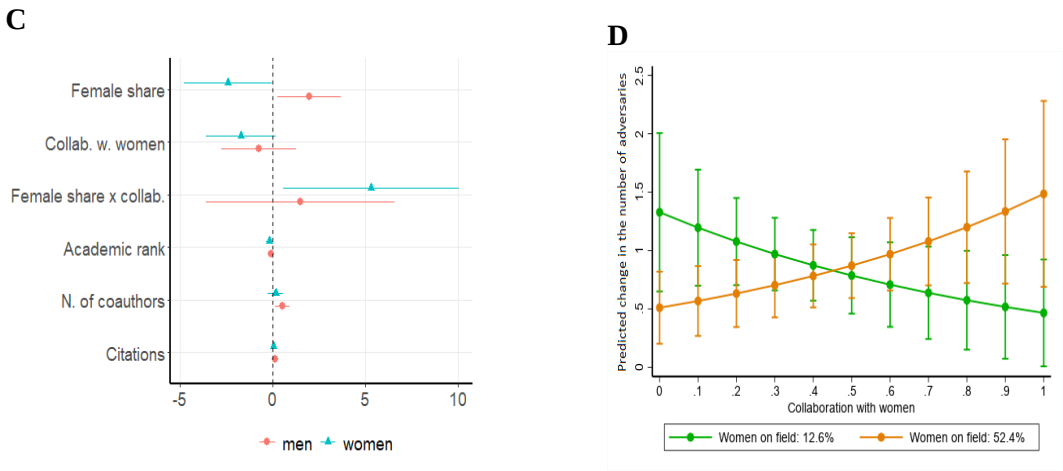


Figure 4. A. Coefficients of scientific fields from the model presented in Figure 3C (with confidence intervals) arranged by the share of women on fields, and two regression lines fitted on the coefficients. B. Results of Poisson regressions (random intercept models) by genders on the number of adversaries considering the share of female academics on the field (coefficients and confidence intervals) C. Results of Poisson regressions (random intercept models) by genders on the number of adversaries considering the share of female academics, the collaboration with women and their interactions. D. Predicted change in the number of adversaries for women by the share of women on the field (lowest and highest observed values) and the individual's collaboration strength with women. Notes. N= 422 men + 301 women. Scale of variables: N. of coauthors (100), Clustering (0.1).

Discussion

The research and pursuit of gender equality now has a strong and colorful tradition (Clavero & Galligan, 2021; Nielsen, 2014; Squires, 2007). In academia, despite all efforts, however, we still observe a huge gender gap. It can be illustrated by the fact, for example, that only 15% of highly cited researchers are women, while 33% of all authors are female (Meho, 2022). Moreover, this gap did not improve at all during the last decade (Lietz et al., 2024; Meho, 2022). While the presence of the gap is evident, research on the mechanisms behind this can easily provoke heated discussions over statistical methodology and on the conclusions one can draw from the analysis - recent examples include e.g. Strumia (2021) or AlShebli et al., (2020). This sensitivity of the question is related to its high policy relevance. While neglecting extant inequalities can be harmful on the one hand, exaggerating difficulties may contribute to stereotypes that can be a force of deterrence for young talents (Ball et al., 2021). Still, several tendencies can be taken as evident. One is the higher dropout rate of women over the career that creates a "leaky pipeline" in their representation over the different ranks (Dubois-Shaik & Fusulier, 2015), and second, the Matthew-effect that success early in the career determines later success versus dropout (Guan et al., 2017).

We attempt to enrich this literature by highlighting a new element, the role of negative ties. Using survey data, we demonstrate that young female academics experience more conflicts than young male scientists do - by large. Furthermore, our results suggest that these conflict relations may create an obstacle in the career of young female scientists in brokering roles between different communities, that are more promising in terms of scientific success (Guan et al., 2017; Jadidi et al., 2018), because in these situations they face more conflicts. Seeing that in our data junior women are more likely to report negative relations than seniors, we may speculate that the gender differences in conflict relations may contribute to the higher dropout rate of young female scientists. Thus, although the same collaboration patterns were found useful for scientists regardless of gender (Dorantes-Gilardi et al., 2023; Jadidi et al., 2018), it seems that the community expects different behavior from men and women Additionally, we do not see that diversity (gender balance of the field) would significantly alter this situation. However, we must also note, that scientific success is not without conflicts for men either; we see that men who are more successful in science (in terms of citations) also report more conflict relations.

What is also interesting is how gender diversity of the field adds to this picture. In this aspect, we see that men working in male-dominated fields report fewer conflicts than men in balanced fields. For women, however, we see interesting interactions. They report more conflicts if they work with men in male-dominated fields, and if they work with women in balanced fields. Given that homophily in collaboration leads to information disadvantage for minority groups (Karimi et al., 2018), this again creates a trade-off for women working in male-dominated fields; they either work with men and risk conflicts or work with each other that may create a lock-in that is less fruitful on the long term. This, however, underlines the importance of initiatives that increase visibility and promote cooperation between women in fields, where they traditionally have low representation; for instance, Women in Data Science, the Society of Women Engineers, and Women in Aviation International. In more diverse fields this trade-off does not exist, as working in gender-integrated teams tends to be both more productive (Vedres & Vásárhelyi, 2023) and also less exposed to conflicts. In this regard, Merluzzi finds that women are more inclined than men to cite a negative relationship with another woman if they lack female social support in the workplace network. Our results align with the assumption that co-authorship is not only a strategic collaboration aiming at career advancement but also a means of finding support in gender-homophile contacts and working together with potential role models (Ely, 1994; Duguid, 2011).

At this point we need to point out a limitation of our study, that is we consider these scientific fields uniform and do not take into account gender segregation within them, across sub-fields (Bandelj, 2019; Strumia, 2021). Considering this, for example, one can imagine a cohesive and gender-homophile female collaboration network in a seemingly male-dominated field too).

References

AlShebli, B., Makovi, K., & Rahwan, T. (2020). RETRACTED ARTICLE: The association between early career informal mentorship in academic collaborations and junior author performance. *Nature Communications*, *11*(1), 1–8.

Aral, S., & Dhillon, P. S. (2023). What (Exactly) Is Novelty in Networks? Unpacking the Vision Advantages of Brokers, Bridges, and Weak Ties. *Management Science*, *69*(2), 1092–1115. https://doi.org/10.1287/mnsc.2022.4377

Ball, P., Britton, T. B., Hengel, E., Moriarty, P., Oliver, R. A., Rippon, G., Saini, A., & Wade, J. (2021). Gender issues in fundamental physics: Strumia's bibliometric analysis fails to account for key confounders and confuses correlation with causation. *Quantitative Science Studies*, *2*(1), 263–272. https://doi.org/10.1162/qss_a_00117

Bandelj, N. (2019). Academic Familism, Spillover Prestige and Gender Segregation in Sociology Subfields: The Trajectory of Economic Sociology. *The American Sociologist*, *50*(4), 488–508. https://doi.org/10.1007/s12108-019-09421-4

Barthauer, L., Spurk, D., & Kauffeld, S. (2016). Women's Social Capital in Academia: A Personal Network Analysis. *International Review of Social Research*, *6*(4), 195–205. https://doi.org/10.1515/irsr-2016-0022

Benan, K. Y., & Olca, S. D. (2020). Cinsiyete Dayalı Tokenizm: Kadın Egemen ve Erkek Egemen Meslekler Üzerinde Nitel Bir Araştırma. *Istanbul Management Journal*, 85–125. https://doi.org/10.26650/imj.2020.88.0004

Bozeman, B., & Gaughan, M. (2011). How do men and women differ in research collaborations? An analysis of the collaborative motives and strategies of academic researchers. *Research Policy*, *40*(10), 1393–1402. https://doi.org/10.1016/j.respol.2011.07.002

Burt, R. S. (1992). *Structural Holes: The Social Structure of Competition*. Harvard University Press. https://www.jstor.org/stable/j.ctv1kz4h78

Burt, R. S. (1998). THE GENDER OF SOCIAL CAPITAL. *Rationality and Society*, *10*(1), 5–46. https://doi.org/10.1177/104346398010001001

Carboni, I. (2023). Women Alone in the Middle: Gender Differences in the Occupation and Leverage of Social Network Brokerage Roles. In A. Gerbasi, C. Emery, & A. Parker (Eds.), *Understanding Workplace Relationships* (pp. 101–134). Springer International Publishing. https://doi.org/10.1007/978-3-031-16640-2_4

Carboni, I., & Gilman, R. (2012). Brokers at risk: Gender differences in the effects of structural position on social stress and life satisfaction. *Group Dynamics: Theory, Research, and Practice*, *16*(3), 218–230. https://doi.org/10.1037/a0028753

Clavero, S., & Galligan, Y. (2021). Delivering gender justice in academia through gender equality plans? Normative and practical challenges. *Gender, Work & Organization*, *28*(3), 1115–1132. https://doi.org/10.1111/gwao.12658

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, *94*, S95–S120.

Coleman, J. S. (1990). *Foundations of social theory*. Belknap Press of Harvard Univ. Press.

Dahlander, L., & McFarland, D. A. (2013). Ties That Last: Tie Formation and Persistence in Research Collaborations over Time. *Administrative Science Quarterly*, *58*(1), 69–110. https://doi.org/10.1177/0001839212474272

Derks, B., Ellemers, N., Van Laar, C., & De Groot, K. (2011). Do sexist organizational cultures create the Queen Bee? *British Journal of Social Psychology*, *50*(3), 519–535. https://doi.org/10.1348/014466610X525280

Dorantes-Gilardi, R., Ramírez-Álvarez, A. A., & Terrazas-Santamaría, D. (2023). Is there a differentiated gender effect of collaboration with super-cited authors? Evidence from junior researchers in economics. *Scientometrics*, *128*(4), 2317–2336. https://doi.org/10.1007/s11192-023-04656-y

Dubois-Shaik, F., & Fusulier, B. (2015). *Academic Careers and Gender Inequality: Leaky Pipeline and Interrelated Phenomena in Seven European Countries*. https://eige.europa.eu/sites/default/files/garcia_working_paper_5_academic_careers_gender_inequality.pdf

Duguid, M. (2011). Female tokens in high-prestige work groups: Catalysts or inhibitors of group diversification? *Organizational Behavior and Human Decision Processes*, *116*(1), 104–115. https://doi.org/10.1016/j.obhdp.2011.05.009

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation* (pp. xii, 178). Lawrence Erlbaum Associates, Inc.

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. https://doi.org/10.1037/0033-295X.109.3.573

Elsesser, K. M., & Lever, J. (2011). Does gender bias against female leaders persist? Quantitative and qualitative data from a large-scale survey. *Human Relations*, *64*(12), 1555–1578. https://doi.org/10.1177/0018726711424323

Ely, R. J. (1994). The Effects of Organizational Demographics and Social Identity on Relationships among Professional Women. *Administrative Science Quarterly*, *39*(2), 203. https://doi.org/10.2307/2393234

Fronczak, A., Mrowinski, M. J., & Fronczak, P. (2022). Scientific success from the perspective of the strength of weak ties. *Scientific Reports*, *12*(1), 5074. https://doi.org/10.1038/s41598-022-09118-8

Gersick, C. J. G., Dutton, J. E., & Bartunek, J. M. (2000). LEARNING FROM ACADEMIA: THE IMPORTANCE OF RELATIONSHIPS IN PROFESSIONAL LIFE. *Academy of Management Journal*, *43*(6), 1026–1044. https://doi.org/10.2307/1556333

Gonzalez-Brambila, C. N. (2014). Social capital in academia. *Scientometrics*, *101*(3), 1609–1625. https://doi.org/10.1007/s11192-014-1424-2

Guan, J., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics*, *11*(2), 407–422. https://doi.org/10.1016/j.joi.2017.02.007

Heilman, M. E., & Okimoto, T. G. (2007). Why are women penalized for success at male tasks?: The implied communality deficit. *Journal of Applied Psychology*, *92*(1), 81–92. https://doi.org/10.1037/0021-9010.92.1.81

Holgersson, C., & Romani, L. (2020). Tokenism Revisited: When Organizational Culture Challenges Masculine Norms, the Experience of Token Is Transformed. *European Management Review*, *17*(3), 649–661. https://doi.org/10.1111/emre.12385

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, *117*(9), 4609–4616. https://doi.org/10.1073/pnas.1914221117

Ibarra, H. (1997). Paving an Alternative Route: Gender Differences in Managerial Networks. *Social Psychology Quarterly*, *60*(1), 91. https://doi.org/10.2307/2787014

Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2018). GENDER DISPARITIES IN SCIENCE? DROPOUT, PRODUCTIVITY, COLLABORATIONS AND SUCCESS OF MALE AND FEMALE COMPUTER SCIENTISTS. *Advances in Complex Systems*, *21*(03n04), 1750011. https://doi.org/10.1142/S0219525917500114

Jansen, D., Von Görtz, R., & Heidler, R. (2010). Knowledge production and the structure of collaboration networks in two scientific fields. *Scientometrics*, *83*(1), 219–241. https://doi.org/10.1007/s11192-009-0022-1

Kane, G., & Labianca, G. (2006). *Accounting for clergy's social ledgers: Mixed blessings associated with direct and indirect negative ties in a religious organization*.

Kanter, R. M. (1977). Some Effects of Proportions on Group Life: Skewed Sex Ratios and Responses to Token Women. *American Journal of Sociology*, *82*(5), 965–990. https://doi.org/10.1086/226425

Karimi, F., Génois, M., Wagner, C., Singer, P., & Strohmaier, M. (2018). Homophily influences ranking of minorities in social networks. *Scientific Reports*, *8*(1), 11077. https://doi.org/10.1038/s41598-018-29405-7

Kwiek, M., & Roszka, W. (2021). Gender-based homophily in research: A large-scale study of man-woman collaboration. *Journal of Informetrics*, *15*(3), 101171. https://doi.org/10.1016/j.joi.2021.101171

Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Bibliometrics: Global gender disparities in science. *Nature*, *504*(7479), 211–213. https://doi.org/10.1038/504211a

Lietz, H., Jadidi, M., Kostic, D., Tsvetkova, M., & Wagner, C. (2024a). Individual and gender inequality in computer science: A career study of cohorts from 1970 to 2000. *Quantitative Science Studies*, *5*(1), 128–152. https://doi.org/10.1162/qss_a_00283

Lietz, H., Jadidi, M., Kostic, D., Tsvetkova, M., & Wagner, C. (2024b). Individual and gender inequality in computer science: A career study of cohorts from 1970 to 2000. *Quantitative Science Studies*, *5*(1), 128–152. https://doi.org/10.1162/qss_a_00283

Lutter, M. (2015). Do Women Suffer from Network Closure? The Moderating Effect of Social Capital on Gender Inequality in a Project-Based Labor Market, 1929 to 2010. *American Sociological Review*, *80*(2), 329–358. https://doi.org/10.1177/0003122414568788

Marsden, P. V. (1990). Network Data and Measurement. *Annual Review of Sociology*, *16*(1), 435–463. https://doi.org/10.1146/annurev.so.16.080190.002251

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Meho, L. I. (2022). Gender gap among highly cited researchers, 2014–2021. *Quantitative Science Studies*, *3*(4), 1003–1023. https://doi.org/10.1162/qss_a_00218

Merluzzi, J. (2017). Gender and Negative Network Ties: Exploring Difficult Work Relationships Within and Across Gender. *Organization Science*, *28*(4), 636–652. https://doi.org/10.1287/orsc.2017.1137

Nahapiet, J., & Ghoshal, S. (1998). Social Capital, Intellectual Capital, and the Organizational Advantage. *The Academy of Management Review*, *23*(2), 242. https://doi.org/10.2307/259373

Newman, M. E. J. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical Review E*, *64*(1), 016132. https://doi.org/10.1103/PhysRevE.64.016132

Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, *45*(2), 167–256. https://doi.org/10.1137/S003614450342480

Nielsen, M. W. (2014). Justifications of Gender Equality in Academia: Comparing Gender Equality Policies of Six Scandinavian Universities. *NORA - Nordic Journal of Feminist and Gender Research*, *22*(3), 187–203. https://doi.org/10.1080/08038740.2014.905490

Onnela, J.-P., Saramäki, J., Kertész, J., & Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, *71*(6), 065103. https://doi.org/10.1103/PhysRevE.71.065103

Patel, V. M., Panzarasa, P., Ashrafian, H., Evans, T. S., Kirresh, A., Sevdalis, N., Darzi, A., & Athanasiou, T. (2019). Collaborative patterns, authorship practices and scientific success in biomedical research: A network analysis. *Journal of the Royal Society of Medicine*, *112*(6), 245–257. https://doi.org/10.1177/0141076819851666

Pelled, L. H. (1996). RELATIONAL DEMOGRAPHY AND PERCEPTIONS OF GROUP CONFLICT AND PERFORMANCE: A FIELD INVESTIGATION. *International Journal of Conflict Management*, *7*(3), 230–246. https://doi.org/10.1108/eb022783

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., & Aral, S. (2022). A causal test of the strength of weak ties. *Science*, *377*(6612), 1304–1310. https://doi.org/10.1126/science.abl4476

Ridgeway, C. L. (2001). Gender, Status, and Leadership. *Journal of Social Issues*, *57*(4), 637–655. https://doi.org/10.1111/0022-4537.00233

Ridgeway, C. L. (2009). Framed Before We Know It: How Gender Shapes Social Relations. *Gender & Society*, *23*(2), 145–160. https://doi.org/10.1177/0891243208330313

Rudman, L. A., & Glick, P. (2001). Prescriptive Gender Stereotypes and Backlash Toward Agentic Women. *Journal of Social Issues*, *57*(4), 743–762. https://doi.org/10.1111/0022-4537.00239

Schoen, C., Rost, K., & Seidl, D. (2018). The influence of gender ratios on academic careers: Combining social networks with tokenism. *PLOS ONE*, *13*(11), e0207337. https://doi.org/10.1371/journal.pone.0207337

Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). SOCIAL NETWORKS AND THE PERFORMANCE OF INDIVIDUALS AND GROUPS. *Academy of Management Journal*, *44*(2), 316–325. https://doi.org/10.2307/3069458

Squires, J. (2007). *The New Politics of Gender Equality*. Bloomsbury Publishing.

Strumia, A. (2021). Gender issues in fundamental physics: A bibliometric analysis. *Quantitative Science Studies*, *2*(1), 225–253. https://doi.org/10.1162/qss_a_00114

Szell, M., & Thurner, S. (2013). How women organize social networks different from men. *Scientific Reports*, *3*(1), 1214. https://doi.org/10.1038/srep01214

Vedres, B., & Vásárhelyi, O. (2023). Inclusion unlocks the creative potential of gender diversity in teams. *Scientific Reports*, *13*(1), 13757. https://doi.org/10.1038/s41598-023-39922-9

Venkataramani, V., Labianca, G. (Joe), & Grosser, T. (2013). Positive and negative workplace relationships, social satisfaction, and organizational attachment. *Journal of Applied Psychology*, *98*(6), 1028–1039. https://doi.org/10.1037/a0034090

YAP, J., & HARRIGAN, N. (2015). Why does Everybody Hate me? Balance, Status, and Homophily: The Triumvirate of Signed Tie Formation. *Social Networks*, *40*, 103–122. https://doi.org/10.1016/j.socnet.2014.08.002

Zeng, X. H. T., Duch, J., Sales-Pardo, M., Moreira, J. A. G., Radicchi, F., Ribeiro, H. V., Woodruff, T. K., & Amaral, L. A. N. (2016). Differences in Collaboration Patterns across Discipline, Career Stage, and Gender. *PLOS Biology*, *14*(11), e1002573. https://doi.org/10.1371/journal.pbio.1002573

Zimmer, L. (1988). Tokenism and Women in the Workplace: The Limits of Gender-Neutral Theory. *Social Problems*, *35*(1), 64–77. https://doi.org/10.2307/800667

https://doi.org/10.51408/issi2025_034

Green or Gold: Exploring How Open Access Models Shape Global Research Integrity

Denis Kosyakov

kosyakov@sciencepulse.ru
Russian Research Institute of Economics, Politics and Law in Science and Technology (RIEPL)

Abstract

Open Access (OA) was conceived to democratize scientific knowledge, yet concerns have arisen about how different OA models affect research integrity. This study examines the relationship between two major publishing pathways – Gold OA and Green OA – and academic integrity across 60 countries and multiple disciplines from 2014 to 2023, drawing on Scopus-indexed journal publications. Gold OA, often operating under a pay-to-publish model, has been criticized for creating incentives that potentially erode the quality of peer review, fostering predatory journals, and disadvantaging authors lacking financial resources. Green OA, on the other hand, allows researchers to self-archive their work, thereby reducing financial barriers and potentially promoting transparency and reproducibility. To gauge research integrity, we use a composite score based on the share of publications in journals that Scopus has discontinued for quality concerns, and the share of retracted articles, giving heavier weight to retractions. Regression analyses reveal a statistically significant negative association between Gold OA share and the transformed integrity score, whereas a higher Green OA share correlates positively with research integrity. However, the explanatory power of these variables is moderate (Adj. $R^2 \approx 0.288$), indicating that other factors also play pivotal roles. Further stratified analyses by discipline show that both Gold and Green OA practices vary by field, but the link between OA model and integrity indicators remains consistent overall: Gold OA tends to correlate with lower integrity, while Green OA is generally associated with higher integrity. National research culture appears to be especially influential, possibly due to varying systems of performance evaluation, career advancement, and ethical oversight. These findings underscore the need for careful policy considerations in promoting OA. While OA can expand accessibility and foster more equitable knowledge dissemination, the manner in which OA is implemented can have unintended consequences for scholarly standards.

Introduction

The Open Access (OA) movement was initially conceived as a mechanism to democratize access to scholarly research. By making publicly funded studies freely accessible, OA aimed to foster greater equity and collaboration within the scientific community. However, in practice, its evolution has raised new questions about research quality and integrity, especially in the context of the pay-to-publish Gold OA model, which some argue has led to the co-option of the movement by commercial interests (Arthur et al., 2023).

Richard Poynder, a noted commentator on scholarly communication (Anderson, 2023; Poynder, 2020) has expressed disappointment that the OA movement has failed to deliver on its promises of accessibility, affordability, and equity. Poynder believes that insufficient advocacy and oversight enabled organizations with different priorities to steer the movement away from its original mission. He further criticizes the pay-to-publish model, contending that it exacerbates affordability problems, marginalizes unfunded researchers and scholars in lower-income regions,

and generally intensifies bureaucratic processes without ensuring meaningful reform.

A key concern regarding the Gold OA model is the proliferation of predatory journals (Beall, 2012). By exploiting author-paid fees, such journals prioritize profit over editorial quality, leading to poor peer review and deceptive practices. This environment can facilitate the publication of low-quality or fraudulent research, eroding trust in scientific publishing. High article processing charges (APCs) in Gold OA also disproportionately affect scholars from under-resourced institutions or countries, independent researchers and pilot studies not supported by research grants, thereby reinforcing global inequities in research dissemination and visibility (Klebel & Ross-Hellauer, 2023).

The pay-to-publish structure of Gold OA creates potential conflicts of interest, where publishers have financial incentives to accept more papers, potentially compromising the peer-review process. Authors, under pressure to publish for career advancement, may be more inclined to submit low-quality or even unethical work. Funders, eager to demonstrate their support for transparency and dissemination, may fail to adequately monitor the integrity of the publications they sponsor. This confluence of interests has led to concerns that Gold OA may inadvertently facilitate research misconduct, including plagiarism, fabrication and falsification, salami slicing of publications, and even an authorship commerce (Chirico & Bramstedt, 2023).

Hanson et al. (Hanson et al., 2024) describe the Gold OA model as "the love triangle of scientific publishing" , in which publishers, authors, and funders are interconnected by financial motivations rather than a unified commitment to scholarly rigor. Publishers benefit from additional article fees, funders rely on publication volume to distribute grants and positions, and researchers need frequent publications to maintain or advance their careers. These interactions drive the growth in scientific publications, often leading to a trade-off between quantity and quality.

Supporters of the traditional subscription model emphasize that university research libraries and their patrons historically served as de facto quality gatekeepers. Librarians, guided by budget constraints and reader feedback, carefully selected reputable journals, thereby curbing the proliferation of low-quality or predatory outlets (Ojennus, 2019). However, this model has been gradually undermined by bundled "big deal" subscriptions offered by major publishers. When libraries must purchase large journal packages rather than selecting titles individually, they lose the granular control essential for maintaining high scholarly standards (Shu et al., 2018).

The Green OA model supports knowledge equity by allowing researchers from diverse backgrounds to access and contribute to scientific knowledge without financial barriers. By enabling self-archiving, Green OA reduces reliance on multinational publishing companies, which often dominate the academic publishing landscape and create inequities in knowledge distribution. The model aligns with the principles of open science, which advocate transparent and accessible research processes. Open science practices, such as preprints and open peer reviews, further support the goals of Green OA by making research outputs available to a wider audience and increasing the accountability of the research process. Green OA encourages the sharing of supplementary materials and data, which enhances the

transparency of research findings. This openness allows other researchers to verify results, conduct replication studies, and build upon existing work, thereby promoting reproducibility and scientific integrity (Winker et al., 2023).

Research misconduct is a pervasive issue in the scientific community, with its prevalence varying significantly across countries and subject areas. In Developing and Emerging Economies, the pressure to publish can lead to unethical practices, such as the sale of authorships and the proliferation of "paper mills" (Vasconez-Gonzalez et al., 2024). The lack of stringent regulatory measures and training in research ethics further exacerbates the issue. In South and East Asia, plagiarism is a common form of misconduct, driven by a lack of training in scientific writing and research ethics, as well as permissive attitudes towards such practices (Rodrigues et al., 2023). But this situation is also prevalent in high-income countries, as evidenced by the increasing rates of retractions due to misconduct in Europe (Freijedo-Farinas et al., 2024; Marco-Cuenca et al., 2021). The prevalence of misconduct varies across disciplines, with fields that are more globalized and research-oriented showing lower instances of plagiarism (Guba & Tsivinskaya, 2024). This suggests that national science culture norms and discipline peculiarities can influence the level of academic integrity (Brooker & Allum, 2024; Fanelli et al., 2015).

Given these complexities, this article investigates the impacts of Gold and Green OA models on research integrity. Through an analysis of publishing structures and disciplinary contexts across multiple countries, the study seeks to clarify how different OA pathways can influence researcher behaviour, quality standards, and the global accessibility of scientific knowledge.

Data and Methods

This research utilizes data from the Scopus database for the period 2014–2023 to analyze the effects of Open Access (OA) publishing models on academic integrity across different countries and subject areas. The study focuses on journal research publications, with the following restrictions: we consider documents of source type "journal" and document types "article", "review", "conference paper", "data paper", and "short survey". Data is aggregated for the top 60 countries by publication output and further divided into second-level subject areas as defined by Scopus All Science Journal Classification (ASJC).

Metrics calculated:

- *Total Number of Documents*: The overall count of journal publications in the selected categories.
- *Number of Gold OA Documents*: The count of documents published under the Gold Open Access model.
- *Number of Green OA Documents*: The count of documents available through Green Open Access.
- *Retracted Articles*: The number of articles marked as retracted in Scopus.
- *Discontinued Journal Publications*: The number of articles published in journals that have been discontinued due to publication concerns or listed on the Scopus Radar for potential issues (as per the Scopus Sources List of December 2024).

We also use the Gross National Income per capita (GNIpc) of countries obtained from World Bank Open Data repository.

Proxy Measure of Academic Misconduct

The main reasons for article retractions in academic journals are often linked to research misconduct. Plagiarism is one of the most common reasons for article retraction, data fabrication (making up data) and falsification (manipulating data or images) also frequently lead to retraction. Duplicate publication, also known as redundant publication, that involves publishing the same or substantially similar work in multiple journals, disputes over authorship, including ghost authorship or inappropriately added/removed authors, can also lead to retractions (Malla & Wani, 2024; Sharma et al., 2023; Valz Gris et al., 2024).

Scopus regularly evaluates and discontinues indexing of journals that no longer meet its quality standards. Two primary reasons for discontinuation are "Publication Concerns" and issues detected by the "Radar" system. Publication Concerns refer to problems related to the quality of editorial practices or other issues that impact a journal's suitability for continued coverage in Scopus (Cortegiani et al., 2020). These concerns may include unfair publication practices, publication of low-quality materials that do not meet scientific criteria, data manipulation, violations of publication ethics, lack of proper peer review, artificial inflation of citations. Publication Concerns can be identified by Scopus itself or flagged by the research community. When legitimate concerns are raised, the journal is added to the re-evaluation program and assessed by the Content Selection & Advisory Board (CSAB) in the year the concern is identified. The Radar system is a data analytics algorithm created by Elsevier Data Scientists to identify journal outlier performance in the Scopus database (*Scopus Content Policy and Selection | Elsevier*, 2024). It runs regularly to check all Scopus journals for unusual patterns and behaviours. Some of the key factors that Radar monitors include rapid and unexplainable changes in the number of articles published, unexplainable shifts in the geographical diversity of authors or affiliations, sudden changes in publication topics compared to the journal's stated aims and scope, abnormal self-citation rates, suspicious editorial policies, consistently low influence metrics. The Radar system is designed to improve continuously by incorporating new examples or signals of potential issues. During the period under review, 2% of research articles in journals indexed in Scopus were published in sources later excluded from indexing and 0.07% were retracted.

To assess the prevalence of academic misconduct, we use a composite indicator based on the share of retracted articles and the share of articles published in discontinued journals. The *Integrity Score* is defined as:

$$Integrity\ Score = 1 - Discontinued\ Share - k \times Retracted\ Share$$

where:

- *Discontinued Share*: The proportion of publications in discontinued journals.
- *Retracted Share*: The proportion of retracted articles.
- k: the weighting factor

The weighting factor for retracted articles reflects their higher significance for the indicator and lower frequency compared to articles in discontinued journals.

Statistical Analysis

The study employs regression analysis to explore the relationship between OA models and academic integrity. The dependent variable is the *Integrity Score*, while the independent variables are:

- *Gold OA Share*: The proportion of publications under the Gold OA model.
- *Green OA Share*: The proportion of publications under the Green OA model.

This regression model allows us to assess how different OA approaches correlate with indicators of research integrity, providing insights into the potential influence of publishing models on academic behavior and misconduct.

Results and Discussion

The descriptive statistics show that *Retracted Share* and *Discontinued Share* are both heavily skewed, whereas *Gold OA Share* and *Green OA Share* exhibit near-normal distributions (Table 1). The mean *Discontinued Share* to the mean *Retracted Share* ratio is 43.3 and we can choose this value for the weighting factor *k*.

Table 1. Descriptive statistics.

Retracted Share		*Discontinued Share*		*Gold OA Share*		*Green OA Share*	
Mean	0.000561	Mean	0.024280	Mean	0.314962	Mean	0.383293
Standard Error	2.7E-05	Standard Error	0.001294	Standard Error	0.003813	Standard Error	0.004606
Median	0.000244	Median	0.005473	Median	0.29834	Median	0.366955
Standard Deviation	0.001086	Standard Deviation	0.052088	Standard Deviation	0.153474	Standard Deviation	0.185393
Sample Variance	1.18E-06	Sample Variance	0.002713	Sample Variance	0.023554	Sample Variance	0.034371
Kurtosis	64.91448	Kurtosis	28.42670	Kurtosis	2.256663	Kurtosis	0.110113
Skewness	6.458127	Skewness	4.621121	Skewness	1.169426	Skewness	0.584524
Range	0.017262	Range	0.531361	Range	0.877351	Range	0.914686
Minimum	0	Minimum	0	Minimum	0.04293	Minimum	0.046343
Maximum	0.017262	Maximum	0.531361	Maximum	0.920281	Maximum	0.961029

The resulting *Integrity Score* is high on average (mean: 0.951), with strong negative skewness indicates a long left tail, meaning many scores are near the maximum value (median: 0.980, skewness: -3.508). Extremely high kurtosis indicates a leptokurtic distribution, with a sharp peak and heavy tail. Correlation analysis (Table 2) shows positive correlation between *Gold OA Share* and *Retracted Share*, *Discontinued Share*, negative correlation with *Green OA Share* and *Retracted Share*, *Discontinued Share*. *Integrity Score* negatively correlates to *Gold OA Share* and positively – with *Green OA Share*.

Table 2. Correlation coefficients.

	Retracted Share	*Discontinued Share*	*Gold OA Share*	*Green OA Share*
Discontinued Share	0.094			
Gold OA Share	0.250	0.050		
Green OA Share	-0.024	-0.332	0.419	
Integrity Score	-0.709	-0.769	-0.196	0.245

To address the non-normal distribution of *Integrity Score* we applied Box-Cox transformation with *lambda* value equal to 14.959. The results of regression analysis with *Transformed Integrity Score* as dependent variable and *Gold and Green OA Shares* as independent variables are presented in Table 3.

Table 3. Correlation coefficients.

Dep. Variable:	Tr. Integrity Score	R-squared:	0.289
Model:	OLS	Adj. R-squared:	0.288
Method:	Least Squares	F-statistic:	329.1
No. Observations:	1620	Prob (F-statistic):	1.24e-120
Df Residuals:	1617	Log-Likelihood:	4306.8
Df Model:	2	AIC:	-8608.
Covariance Type:	nonrobust	BIC:	-8591.

	coef	std err	t	P-value	[0.025	0.975]
Intercept	-0.02849	0.001114	-25.5791	1.83E-121	-0.0307	-0.0263
Gold OA Share	-0.05856	0.003026	-19.3518	3.27E-75	-0.0645	-0.0526
Green OA Share	0.05862	0.002505	23.4036	1.46E-104	0.0537	0.0635

Omnibus:	12.745	Durbin-Watson:	1.530
Prob(Omnibus):	0.002	Jarque-Bera (JB):	12.617
Skew:	-0.196	Prob(JB):	0.00182
Kurtosis:	2.817	Cond. No.	8.91

Approximately 28.9% of the variation in the *Transformed Integrity Score* is explained by the independent variables (*Gold and Green OA Shares*). While this is a moderate level of explanatory power, it suggests other unobserved factors are influencing the integrity score. Adjusted R-squared indicates that the model's explanatory power is robust and not overfitted. The overall model is statistically significant, meaning *Gold OA* and *Green OA* collectively explain significant variation in the dependent variable. Results of Omnibus, Jarque-Bera statistical tests indicate that the residuals deviate slightly negatively from normality, Durbin-Watson

test indicates no significant autocorrelation in residuals and Condition Number indicates no significant multicollinearity issues among predictors.

There is a statistically significant negative relationship between *Gold OA share* and the *Transformed Integrity Score* and positive relationship between *Green OA share* and the *Transformed Integrity Score*. We can assume that higher *Gold OA share* correlates with lower integrity, potentially reflecting issues such as predatory publishing or compromised peer review while higher *Green OA share* correlates with better academic integrity, aligning with the idea that Green OA promotes transparency and good research practices.

Among the articles in journals excluded from indexing, a slightly larger share is accounted for by Gold OA journals - 24.5%, 17.2% are articles in Green OA. Of the retracted articles, 44.8% are from Gold OA, 36.6% are from Green OA. Overall, Green OA accounted for 31.8% of papers out of 25.7 million scientific journal publications from 2014-2023, Gold OA accounted for 24.8%. The correlation coefficient between *Gold OA Share* and *Retracted Share* is 0.25, between *Gold OA Share* and *Discontinued Share* is 0.05. This may to some extent account for the detected correlation but does not explain it completely.

The analysis by fields of science generally shows no difference in the correlation between Green and Gold OA Shares and Integrity Score. In both cases, a weak negative correlation is observed, i.e. a larger share of documents in any type of OA is more likely to correspond to a higher level of academic integrity. At the same time, disciplinary specificity is present, both in open access practices and, presumably, in the manifestations of questionable research practices leading to retraction of articles and exclusion of journals from indexing. If we look at scientific fields in the context of national segments of science, a negative correlation also prevails in both cases: in 42 out of 60 countries. It is worth noting that in 27 cases the negative correlation of *Research Integrity Score* with *Gold OA Share* is more pronounced than with *Green OA Share*. In three cases, *Research Integrity Score* is negatively correlated with *Gold OA Share*, while positively correlated with *Green OA Share*; in one case, the opposite is true; in 14 cases, the correlation with *Green OA Share* is negative.

For countries in general, the difference in the dependence of *Research Integrity Score* on the share of articles in Gold and Green OA is clearly visible (Fig. 1), demonstrating the significant influence of the specifics of the national research environment.

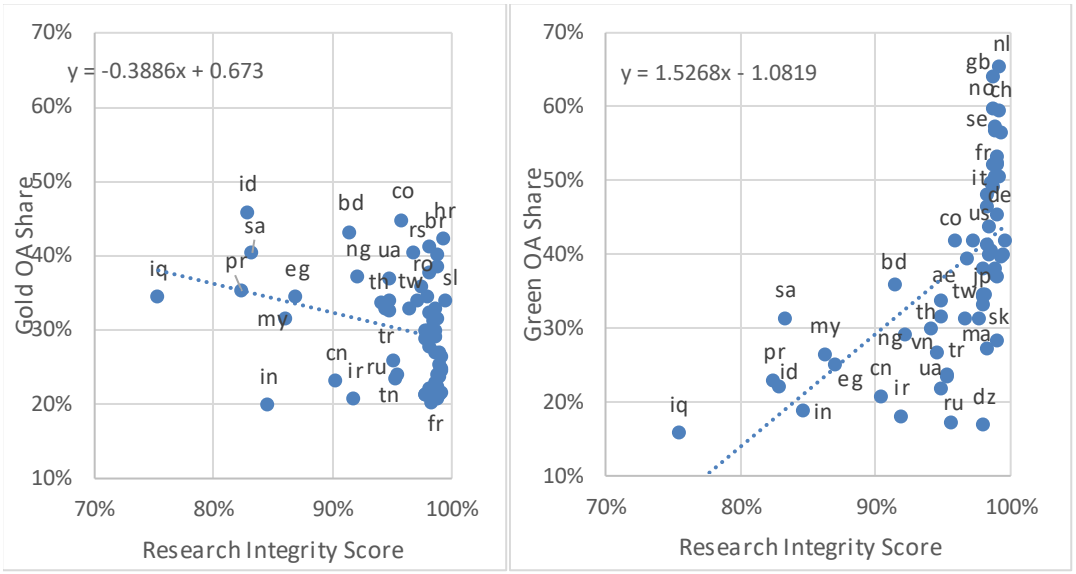


Figure 1. Correlation between OA Share and Research Integrity Score for different countries.

This value is also observed in individual research areas, a higher share of Green OA in a country corresponds to a higher *Integrity Score*, while a higher share of Gold OA, on the contrary, is associated with lower *Integrity Score* values in 24 out of 27 areas. The exceptions are *Multidisciplinary*, *Physics and Astronomy*, *and Environmental Science*. Fig. 2 shows the research area of *Business, Management and Accounting*.

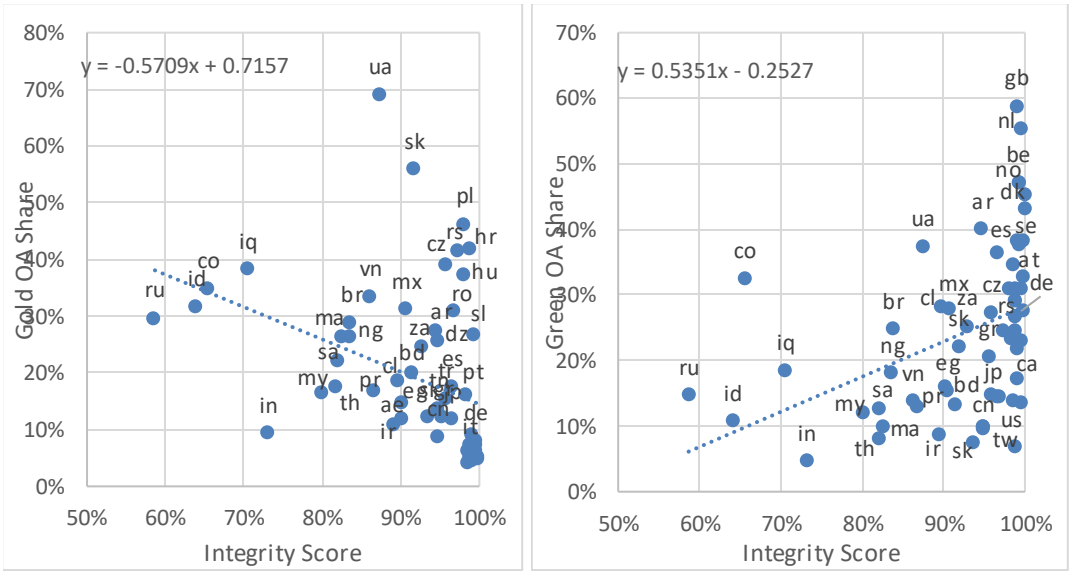


Figure 2. Correlation between OA Share and Research Integrity Score for different countries in the Business, Management and Accounting research area.

A striking contrast emerges when comparing the two countries (Fig. 3). In Indonesia, a developing economy, Green OA initially shows modest gains but soon stagnates

and even declines, whereas Gold OA experiences rapid growth, eventually surpassing Green OA by a wide margin. This pattern suggests that authors in Indonesia may be gravitating toward pay-to-publish outlets – possibly due to perceptions of prestige or the lack of robust institutional repositories – leading to a smaller share of self-archived content.

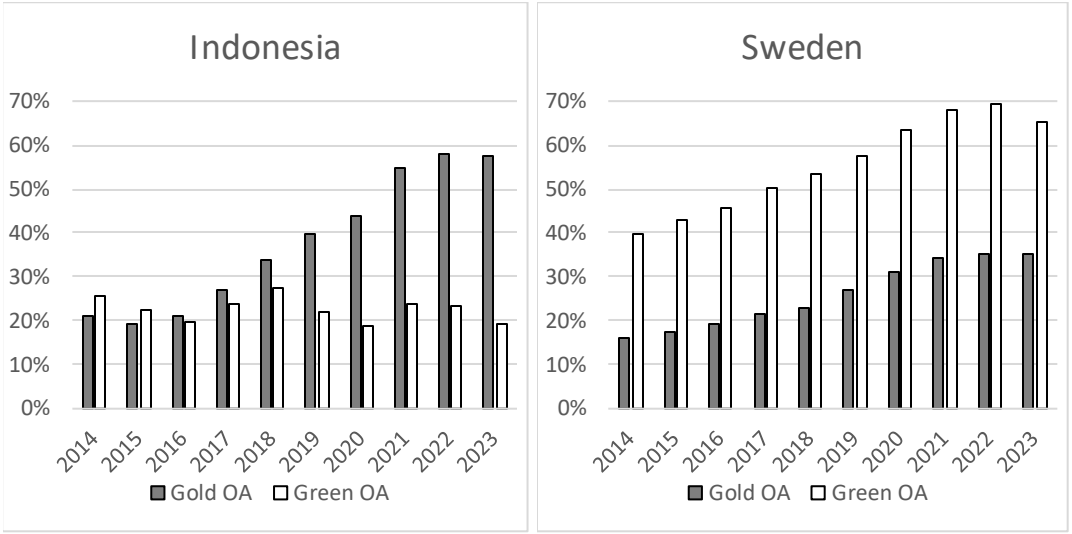


Figure 3. Correlation between OA Share and Research Integrity Score for different countries in the Business, Management and Accounting research area.

By contrast, Sweden's moderate but steady increase in Gold OA coexists with a high and growing proportion of Green OA publications. This can be partly attributed to institutional mandates and research funders' requirements, which encourage or even oblige Swedish researchers to deposit their work in open repositories. Such policies offer a sustainable, non-commercial pathway to openness and thus maintain a strong Green OA presence while still allowing for a measured growth in Gold OA. This is a typical picture reflecting the situation in developed and developing countries.

We propose an indicator characterizing the difference in document shares between Green OA and Gold OA. For several countries, this ratio is negative, indicating that the share of Gold OA publications consistently exceeds that of Green OA. Notably, most of these countries also exhibit relatively low *Research Integrity Scores*. In contrast, countries where Green OA predominates tend to have higher *Research Integrity Scores*, with a correlation coefficient of 0.64 (Fig. 4).

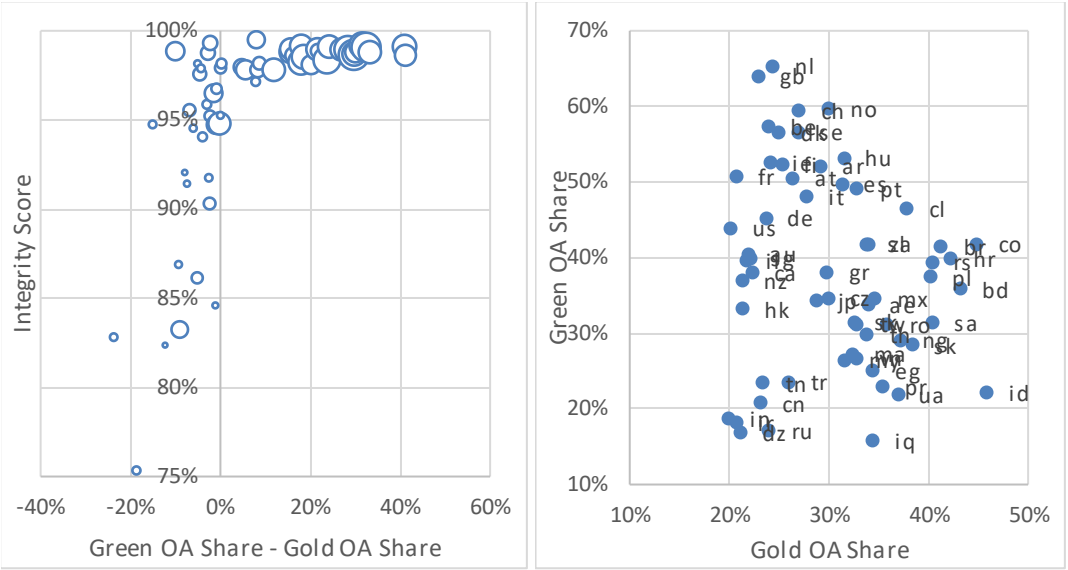


Figure 4. Correlation between *Research Integrity Score* and difference between *Gold OA Share* and *Green OA Share* (left, size of the bubble corresponds to GNIpc); *Gold OA Share* and *Green OA Share* (right) for different countries.

At the disciplinary level, no significant correlation is observed between these indicators, as evidenced by a correlation coefficient of -0.02. When examining the correlation at the country-discipline level, which was used to construct the regression model, the correlation coefficient is slightly lower, at 0.41.

The findings support the hypothesis that a high proportion of publications in Gold OA is associated with a greater prevalence of questionable research practices related to violations of research integrity. This relationship is further exacerbated when the share of Green OA publications is low. Moreover, the prevalence of questionable research practices appears to have a strong national component, likely influenced by variations in national research cultures. These variations are shaped by differing levels of publication pressure, which may result from policies on the certification of scientific personnel, incentives for publication activity, and mechanisms for evaluating scientific performance. It should be noted that the observed dependence is influenced by disciplinary characteristics, which can probably offset the impact of national research culture.

Limitations

While our study provides meaningful insights into the relationship between Open Access (OA) models and research integrity, several limitations should be noted.

First, the Integrity Score used in our analysis is a composite indicator based on the proportions of retracted articles and publications in discontinued journals. Although article retractions typically indicate serious misconduct such as plagiarism, data fabrication, or falsification (Malla & Wani, 2024; Sharma et al., 2023; Valz Gris et al., 2024), not all retractions necessarily reflect intentional misconduct; some result from honest errors or disputes unrelated to ethical breaches. Similarly, journal

discontinuations, as explained previously, can occur due to various quality-related issues identified by Scopus, including editorial misconduct, poor peer review practices, or abnormal citation patterns. Thus, the Integrity Score should be considered indicative rather than definitive.

Second, although our model identified statistically significant relationships, the moderate explanatory power suggests that other relevant factors influencing research integrity were not captured in this study. Variables such as funding mechanisms, institutional policies, individual researcher motivations, or detailed disciplinary cultures could substantially affect research integrity, warranting further exploration.

Third, while we highlighted the role of national research cultures, our study does not operationalize this variable quantitatively. A systematic characterization, possibly incorporating data from worldwide surveys, could provide deeper insights into cultural determinants of research integrity.

Finally, our analysis does not deeply investigate disciplinary differences in OA publishing patterns and integrity. Further field-specific analysis could elucidate why disciplines vary in their engagement with different OA models and the resulting implications for research integrity.

Despite these limitations, our findings contribute valuable insights into the ongoing discourse on OA publishing and offer practical policy implications to promote ethical scholarly communication.

Conclusion

Our findings highlight the complex relationship between Open Access (OA) models and research integrity, revealing both opportunities and challenges associated with different publishing approaches. While OA is fundamental to expanding the accessibility of scientific knowledge, its implementation can have divergent consequences. The Gold OA model, which operates on an author-pays principle, exhibits a moderate but consistent negative correlation with research integrity indicators. This association likely reflects the proliferation of predatory publishing practices and the shortcomings of peer review in certain venues that prioritize financial transactions over rigorous editorial standards. This observed correlation does not imply direct causality, as other factors, including publication pressures and weak regulatory frameworks, could simultaneously influence both OA preferences and integrity outcomes. In contrast, Green OA is positively associated with research integrity, reflecting its ability to enhance transparency and reduce financial barriers, thereby supporting more robust ethical practices.

Beyond the specific impact of OA models, our study highlights the decisive role of national research cultures in mediating these effects. Countries with strong regulatory oversight, well-balanced research evaluation systems, and established ethical frameworks appear better equipped to leverage the advantages of OA while minimizing its risks. Conversely, in regions where publication pressure is intense and regulatory mechanisms remain weak, the structural vulnerabilities of the Gold OA model may intensify unethical research practices, including compromised peer review, citation manipulation, and the emergence of low-quality publications. In many developing scientific systems experiencing rapid expansion, the rise of new

research groups and disciplines has outpaced the establishment of a mature research culture. This misalignment fosters an environment in which publication quantity is prioritized over quality, further reinforcing problematic publishing behaviors.

At the same time, it is important to recognize that the accumulation of research culture within emerging scientific communities may gradually improve research integrity over time. Fields that have historically matured within these national systems appear to have already adopted more rigorous ethical standards, demonstrating that research integrity is not inherently constrained by geography or economic conditions but rather by the broader scientific environment in which scholars operate. However, the Gold OA model, due to its inherent conflict of interest where publishers profit directly from article processing charges introduces additional ethical risks, particularly in environments with underdeveloped research cultures. The financial barriers posed by high APCs in leading OA journals may also push researchers from lower-income countries toward lower-ranked or less scrupulous publishing outlets, further intensifying disparities in research quality (Björk & Solomon, 2015).

In addition to these ethical concerns, the Gold OA model imposes a significant financial burden on national R&D sectors, a challenge that is particularly acute in developing economies (Haustein et al., 2024). The substantial funds allocated to cover APCs could be more effectively invested in fostering a more sustainable and ethically robust model of scholarly publishing, such as Diamond OA. Unlike Gold OA, the Diamond OA model removes financial barriers for both authors and readers, offering a more equitable and transparent approach to disseminating research. Redirecting resources toward such initiatives would not only alleviate financial pressures but also contribute to strengthening the overall integrity of scientific publishing by eliminating economic incentives that may encourage questionable research practices (Fuchs & Sandoval, 2013).

These observations lend further support to Poynder's critique that the OA movement has deviated from its original vision. The Budapest Open Access Initiative (*Budapest Open Access Initiative*, 2002), which set the foundation for OA principles, emphasized two complementary strategies: the development of open repositories for self-archiving and the creation of alternative OA journals supported by non-commercial funding models. The declaration envisioned funding sources primarily from research institutions, government agencies, philanthropic donations, and reallocation of resources from discontinued subscription-based journals. Researcher-funded publication, which defines the contemporary Gold OA model, was considered only as a last resort. The current dominance of the author-pays model represents a fundamental departure from these initial ideals, raising concerns about its unintended consequences for research integrity.

While a transition to Green OA alone may not be sufficient to resolve integrity challenges in research communities where questionable practices are prevalent, it is plausible that reducing reliance on Gold OA could help mitigate some of its more problematic effects. The removal of financial incentives that drive ethically dubious publishing behavior, combined with policies promoting open science practices, could accelerate the development of more robust research cultures. In this context,

strengthening institutional repositories and fostering collaborative models of scholarly communication may represent a more sustainable path toward ensuring both accessibility and integrity in scientific publishing.

References

Anderson, R. (2023, December 7). *Where Did the Open Access Movement Go Wrong?: An Interview with Richard Poynder*. The Scholarly Kitchen. https://scholarlykitchen.sspnet.org/2023/12/07/where-did-the-open-access-movement-go-wrong-an-interview-with-richard-poynder/

Arthur, P. L., Hearn, L., Ryan, J. C., Menon, N., & Khumalo, L. (2023). Making Open Scholarship More Equitable and Inclusive. *Publications*, *11*(3), 41. https://doi.org/10.3390/publications11030041

Beall, J. (2012). Predatory publishers are corrupting open access. *Nature*, *489*(7415), 179–179. https://doi.org/10.1038/489179a

Björk, B.-C., & Solomon, D. (2015). Article processing charges in OA journals: Relationship between price and quality. *Scientometrics*, *103*(2), 373–385. https://doi.org/10.1007/s11192-015-1556-z

Brooker, R., & Allum, N. (2024). Investigating the links between questionable research practices, scientific norms and organisational culture. *Research Integrity and Peer Review*, *9*(1), 12. https://doi.org/10.1186/s41073-024-00151-x

Budapest Open Access Initiative. (2002). https://www.budapestopenaccessinitiative.org/read/

Chirico, F., & Bramstedt, K. A. (2023). Authorship commerce: Bylines for sale. *Accountability in Research*, *30*(4), 246–251. https://doi.org/10.1080/08989621.2021.1982705

Cortegiani, A., Ippolito, M., Ingoglia, G., Manca, A., Cugusi, L., Severin, A., Strinzel, M., Panzarella, V., Campisi, G., Manoj, L., Gregoretti, C., Einav, S., Moher, D., & Giarratano, A. (2020). Citations and metrics of journals discontinued from Scopus for publication concerns: The GhoS(t)copus Project. *F1000Research*, *9*, 415. https://doi.org/10.12688/f1000research.23847.2

Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct Policies, Academic Culture and Career Stage, Not Gender or Pressures to Publish, Affect Scientific Integrity. *PLOS ONE*, *10*(6), e0127556. https://doi.org/10.1371/journal.pone.0127556

Freijedo-Farinas, F., Ruano-Ravina, A., Pérez-Ríos, M., Ross, J., & Candal-Pedreira, C. (2024). Biomedical retractions due to misconduct in Europe: Characterization and trends in the last 20 years. *Scientometrics*, *129*(5), 2867–2882. https://doi.org/10.1007/s11192-024-04992-7

Fuchs, C., & Sandoval, M. (2013). The Diamond Model of Open Access Publishing: Why Policy Makers, Scholars, Universities, Libraries, Labour Unions and the Publishing World Need to Take Non-Commercial, Non-Profit Open Access Serious. *tripleC: Communication, Capitalism & Critique. Open Access Journal for a Global Sustainable Information Society*, *11*(2), 428–443. https://doi.org/10.31269/triplec.v11i2.502

Guba, K. S., & Tsivinskaya, A. O. (2024). Ambiguity in Ethical Standards: Global Versus Local Science in Explaining Academic Plagiarism. *Science and Engineering Ethics*, *30*(1), 4. https://doi.org/10.1007/s11948-024-00464-6

Hanson, M. A., Barreiro, P. G., Crosetto, P., & Brockington, D. (2024). The strain on scientific publishing. *Quantitative Science Studies*, *5*(4), 823–843. https://doi.org/10.1162/qss_a_00327

Haustein, S., Schares, E., Alperin, J. P., Hare, M., Butler, L.-A., & Schönfelder, N. (2024). *Estimating global article processing charges paid to six publishers for open access between 2019 and 2023* (No. arXiv:2407.16551). arXiv. https://doi.org/10.48550/arXiv.2407.16551

Klebel, T., & Ross-Hellauer, T. (2023). The APC-barrier and its effect on stratification in open access publishing. *Quantitative Science Studies*, *4*(1), 22–43. https://doi.org/10.1162/qss_a_00245

Malla, R. A., & Wani, Z. A. (2024). Uncovering the reasons of retraction in virology: A citation and Altmetric investigation. *Global Knowledge, Memory and Communication*, *ahead-of-print*(ahead-of-print). https://doi.org/10.1108/GKMC-11-2023-0415

Marco-Cuenca, G., Salvador-Oliván, J. A., & Arquero-Avilés, R. (2021). Fraud in scientific publications in the European Union. An analysis through their retractions. *Scientometrics*, *126*(6), 5143–5164. https://doi.org/10.1007/s11192-021-03977-0

Ojennus, P. (2019). Modelling advances in gatekeeping theory for academic libraries. *Journal of Documentation*, *76*(2), 389–408. https://doi.org/10.1108/JD-03-2019-0051

Poynder, R. (2020). *Open Access: "Information Wants to Be Free"?* Richard Poynder, Open & Shut? https://digitalcommons.unl.edu/scholcom/182

Rodrigues, F., Gupta, P., Khan, A. P., Chatterjee, T., Sandhu, N. K., & Gupta, L. (2023). The Cultural Context of Plagiarism and Research Misconduct in the Asian Region. *Journal of Korean Medical Science*, *38*(12), e88. https://doi.org/10.3346/jkms.2023.38.e88

Scopus content policy and selection | Elsevier. (2024). Www.Elsevier.Com. https://www.elsevier.com/products/scopus/content/content-policy-and-selection

Sharma, P., Sharma, B., Reza, A., Inampudi, K. K., & Dhamija, R. K. (2023). A systematic review of retractions in biomedical research publications: Reasons for retractions and their citations in Indian affiliations. *Humanities and Social Sciences Communications*, *10*(1), 1–12. https://doi.org/10.1057/s41599-023-02095-x

Shu, F., Mongeon, P., Haustein, S., Siler, K., Alperin, J., & Larivière, V. (2018). Is It Such a Big Deal? On the Cost of Journal Use in the Digital Era. *College & Research Libraries*, *79*(6), 785–798. https://doi.org/10.5860/crl.79.6.785

Valz Gris, A., Cristiano, A., & Pezzullo, A. M. (2024). Integrity and accountability in academic publishing: Trends and implications of paper retractions and journal delistings. *European Journal of Public Health*, *34*(Supplement_3), ckae144.678. https://doi.org/10.1093/eurpub/ckae144.678

Vasconez-Gonzalez, J., Izquierdo-Condoy, J. S., Naranjo-Lara, P., Garcia-Bereguiain, M. Á., & Ortiz-Prado, E. (2024). Integrity at stake: Confronting "publish or perish" in the developing world and emerging economies. *Frontiers in Medicine*, *11*, 1405424. https://doi.org/10.3389/fmed.2024.1405424

Winker, M. A., Bloom, T., Onie, S., & Tumwine, J. (2023). Equity, transparency, and accountability: Open science for the 21st century. *The Lancet*, *402*(10409), 1206–1209. https://doi.org/10.1016/S0140-6736(23)01575-1

https://doi.org/10.51408/issi2025_069

Guidance List for Reporting Bibliometric Analyses (GLOBAL): A Two-Round Modified Delphi Study

Jeremy Y. Ng[1], Henry Liu[1], Mehvish Masood[1], Niveen Syed[1], Ludo Waltman[2], Stefanie Haustein[3,4,5], Michel Sabé[6,7], Marco Solmi[8,9,10,11,12], Dimity Stephen[13]✉, Alexander Schniedermann[13], Simon Willemin[14], Marianne Gauffriau[15], Jens Peter Andersen[16], Lutz Bornmann[17], Rodrigo Costas[2], Wolfgang Glänzel[18], Sybille Hinze[19], Nicolas Robinson-Garcia[20], Gunnar Sivertsen[21], Stephan Stahlschmidt[13], Thed van Leeuwen[2], Erija Yan[22], Cameron Neylon[23], Verena Weimer[24], Alesia Zuccala[25], David Moher [1,26]

[1] Centre for Journalology, Ottawa Methods Centre, Ottawa Hospital Research Institute, Ottawa (Canada)
[2] Centre for Science and Technology Studies, Leiden University, Leiden (The Netherlands)
[3] School of Information Studies, University of Ottawa, Ottawa (Canada)
[4] Scholarly Communications Lab, Ottawa (Canada)
[5] Centre Interuniversitaire de Recherche sur la Science et la Technologie, Université du Québec à Montréal, Montreal (Canada)
[6] Division of Adult Psychiatry, Department of Psychiatry, University Hospitals of Geneva, Thonex (Switzerland)
[7] Faculty of Medicine, University of Geneva, Geneva (Switzerland)
[8] Department of Psychiatry, University of Ottawa, Ottawa (Canada)
[9] Department of Mental Health, The Ottawa Hospital, Ottawa (Canada)
[10] Clinical Epidemiology Program, Ottawa Hospital Research Institute, University of Ottawa, Ottawa (Canada)
[11] School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa (Canada)
[12] Department of Child and Adolescent Psychiatry, Charité Universitätsmedizin, Berlin (Germany)
[13] German Centre for Higher Education Research and Science Studies (DZHW), Berlin (Germany)
[14] ETH Zurich, Zurich (Switzerland)
[15] IT University of Copenhagen, Copenhagen (Denmark)
[16] Danish Centre for Studies in Research and Research Policy, Aarhus University, Aarhus (Denmark)
[17] Administrative Headquarter of the Max Planck Society, Munich (Germany)
[18] KU Leuven, Leuven (Belgium)
[19] Center for Open and Responsible Research, Berlin University Alliance, Berlin (Germany)
[20] University of Granada, Granada (Spain)
[21] Nordic Institute for Studies in Innovation, Research and Education, Oslo (Norway)
[22] Drexel University, Philadelphia (United States of America)
[23] Curtin University, Perth (Australia)
[24] DIPF | Leibniz Institute for Research and Information in Education, Frankfurt am Main (Germany)
[25] University of Copenhagen, Copenhagen (Denmark)
[26] Institute of Health Policy, Management & Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto (Canada)

Abstract

Background: Despite the growth of bibliometric analyses in the scholarly literature, few studies offer guidance on how to report them, resulting in a lack of transparency and completeness in research. To address this gap in thorough reporting practices, in accordance with existing best practice for establishing reporting guidelines, we developed the Guidance List for the repOrting of Bibliometric AnaLyses (GLOBAL), a reporting guideline aimed at promoting high-quality reporting of bibliometric analyses. Methods: An initial list of items for the GLOBAL was generated through a scoping review and further refined through a two-round Delphi, as outlined by the EQUATOR Network's methodological framework on creating reporting guidelines. Participants, including international bibliometric experts, were recruited for the Delphi via personalized emails and open invitations. Consensus was achieved when at least 80% of participants agreed on the inclusion or exclusion of items in the GLOBAL checklist. Items that did not reach consensus were excluded. Round 1, conducted through an international online survey, used a 9-point Likert scale to assess how essential an item was for reporting bibliometric analyses. A content analysis was performed on participant feedback from Round 1, including comments on each item and responses to the open-ended questions. Round 2 consisted of an in-person meeting to discuss and vote on items that were new or did not reach consensus in Round 1. Results: In Round 1, 24 of 32 items reached consensus and content analysis resulted in one new item. This item and the eight items that did not reach consensus were discussed in Round 2. During the meeting, one item was split into two, totalling ten items. Nine out of ten items reached consensus, five for inclusion and four for exclusion, while 1 item was also excluded because it did not reach consensus. Conclusions: The finalized 29-item GLOBAL checklist provides users with guidance to report bibliometric analyses. Its international adoption is aimed at improving the reporting practices of bibliometric analyses for research purposes.

Introduction

Bibliometrics is a social science discipline historically based on three developments: (1) the positivist-functionalist philosophy (of science) of being able to examine social facts objectively; (2) the development of citation indices and analysis to measure research performance; and (3) the discovery of mathematical laws that enabled the use of indicators in science evaluation (De Bellis, 2014). Here, we follow a pragmatic definition of bibliometrics based on common usage in the literature. We define bibliometric analyses as any study that quantitatively studies academic research based on at least one of two basic elements: (1) publications (e.g., journal articles, conference proceedings papers, books and book chapters, preprints, peer review reports, grey literature) to represent scholarly outputs; and (2) citations (i.e., formal references to a publication in the reference lists of other publications) to reflect connections between and the impact of publications. These units of measurement can be applied to various levels of aggregation, for instance: micro-level (e.g., authors, documents), meso-level (e.g., institutions, departments, journals), and macro-level (e.g., countries, disciplines).

Bibliometric analyses may introduce, adapt, and/or apply various types of bibliometric indicators – ranging from absolute numbers of publications and citation rates (e.g., journal impact factor, field-normalized citation rate), citation percentiles, or collaboration strength – to measure, compare and benchmark (AlRyalat et al., 2019; Donthu et al., 2021; Sugimoto & Larivière, 2018). Researchers and organizations conduct bibliometric analyses for a variety of purposes, such as to explore the intellectual structure of an existing field and to identify publication-related characteristics, trends, and patterns specific to a journal, article, book, author,

institution, and/or topic of study. The value of this method is that it enables researchers to discover patterns and "make sense" of a high volume of different characteristics taken from hundreds, thousands, or even millions of publications. The findings of bibliometric analyses can therefore serve to advance a field by providing a comprehensive overview of the research conducted, understanding how research has evolved over time, identifying knowledge gaps, and inspiring novel ideas for investigation in that particular area (Donthu et al., 2021).

Recently, a number of articles have been published describing how to report or conduct a bibliometric analysis (Donthu et al., 2021; Linnenluecke et al., 2020). However, most of these articles have not framed their work in the format of a reporting guideline (Jappe, 2020). The current lack of evidence-based guidance on how to report a bibliometric analysis can be problematic for several reasons. If authors fail to provide readers with enough information about how and when their study was conducted, including e.g., the database from which the bibliographic data were retrieved, readers will only have a partial understanding of what was done. Consequently, insufficient reporting may hinder the reproducibility of a study and further inhibit researchers from evaluating the accuracy of its findings (Bornmann et al., 2021). Furthermore, editors and peer reviewers have no guidelines against which to compare the reporting quality of a study under their consideration. Moral and ethical justifications also exist for providing accurate research reporting (Moher, 2007). Ethical research promotes knowledge, truth, and the avoidance of error, which are values that are essential to both collaborative work and accountability to the public (Resnik, 2015).

As a first step to address this knowledge gap, we opted to develop a reporting guideline for bibliometric analyses, known as the Guidance List for the repOrting of Bibliometric AnaLyses (GLOBAL). A reporting guideline is defined as "a checklist, flow diagram, or explicit text to guide authors in reporting a specific type of research, developed using explicit methodology" (Moher et al., 2010 p. 1). This work stems from our understanding that "bibliometrics" is generally regarded as the most commonly used term, which captures the entire field of research and application that deals with the quantitative analysis of scholarly outputs and their influence.

As bibliometric analyses are increasingly adopted, establishing reporting guidelines for these studies is crucial to strengthening their reliability and accuracy. Such guidelines enhance reporting quality by enabling researchers to ensure their published papers are complete and transparent, thereby positively influencing how researchers plan, execute, and report their work (Donthu et al., 2021; Gagnier et al., 2013; Moher et al., 2010). The GLOBAL has the potential to benefit many stakeholders. As a reporting guideline, the GLOBAL aims to assist researchers in reporting and peer reviewers in evaluating bibliometric analyses. Thorough reporting, supported by adherence to reporting guidelines, allows readers to evaluate the usefulness of a study's methods and, consequently, the reliability and robustness of its conclusions. High-quality reporting may help to ensure new research is efficiently used, less research waste is produced, and may also facilitate easier replication and potential review updates (Moher et al., 2010).

Methods

Study design

The following section briefly outlines the study design, while detailed explanations are provided in the subsequent sections. The GLOBAL was developed in accordance with the EQUATOR Network's methodological framework (EQUATOR Network, n.d.; Moher et al., 2010). A scoping review was conducted to identify relevant reporting guidance for bibliometric analyses and generate a preliminary list of candidate items for the GLOBAL checklist. This scoping review has since been posted as a preprint (Ng et al., 2024). The preliminary list of GLOBAL candidate items was further developed using a two-round modified Delphi that was conducted on a global scale. The Delphi modification came from generating a preliminary list of items through a scoping review and discussions with the GLOBAL steering committee, rather than deriving original ideas from the Delphi panel, although participants could suggest new items during these rounds.

Round 1 of the Delphi involved individuals completing an online survey using Welphi (*Welphi*, n.d.), a web-based platform that is specifically designed to host surveys employing the Delphi method. Round 2 consisted of an in-person consensus group meeting with participants who completed the previous round and were interested and able to attend this meeting. The GLOBAL steering group, which supervised and provided input to the GLOBAL's development, consisted of five international researchers, four with expertise in bibliometrics (LW, MSabé, MSolmi, and SH) and one with expertise in reporting guidelines (DM).

Open science statement

The GLOBAL is registered on the EQUATOR Network Library of Reporting Guidelines (EQUATOR Network, n.d.a). The protocol was registered on January 12, 2023, on the Open Science Framework (OSF) (Ng et al., 2023). Anonymized, aggregate voting data and participant responses from Rounds 1 and 2 were also shared publicly using OSF. Participants in both Delphi survey rounds provided consent to participate in this study. We followed the Accurate Consensus Reporting Document checklist (Gattrell et al., 2024) in reporting our findings.

Scoping review and candidate item generation

An initial list of candidate items for the GLOBAL checklist was generated through a scoping review (Peters et al., 2020) of peer-reviewed literature, articles on preprint servers, and grey literature that aimed to identify and categorize bibliometric reporting recommendations (Ng et al., 2024). Twenty-three studies met the inclusion criteria following screening. Consensus on the inclusion, the section the item belongs to (i.e., 'title', 'abstract', 'introduction', 'methods', 'results', 'discussion', or 'other' sections of the reporting guideline), and the phrasing of candidate items for the GLOBAL were decided after multiple discussions with the steering committee and research team (JYN, HL, MM, NS, LW, MSabé, MSolmi, SH, DS, DM). The steering committee also had the opportunity to add items that seemed necessary to increase the quality of bibliometric reporting but were not addressed by the included

studies (Ng et al., 2024). This process resulted in a 32-item preliminary checklist; 31 items being created based on recommendations from the literature and one item arising from expert opinion of the GLOBAL steering committee.

Recruitment of Delphi Participants

Participants were recruited from a diverse group of international stakeholders with bibliometric experience (e.g., bibliometricians, librarians, journal editors, policy and research analysts, and researchers) through purposeful sampling. Steering committee members did not serve as participants in either Delphi round. Recruitment was conducted through two methods. First, the steering committee compiled a list of experts from the bibliometric community and sent personalized email invitations and reminders to these potential participants through the Welphi platform (*Welphi*, n.d.). Second, an advertisement and recruitment script with a general universal link to the Welphi survey was disseminated to members of the International Society for Scientometrics and Informetrics (ISSI) through their mailing list, an ISSI website blog post on 26 July 2024 and promoted via social media (Twitter, LinkedIn). Information on the GLOBAL Delphi was also listed on the website for the 2024 International Science, Technology and Innovation Indicators (STI) conference website (GLOBAL Delphi Survey, n.d.). The ISSI and STI are communities of researchers and professionals involved in the fields of scientometrics, informetrics, and webometrics. The survey link for both methods (i.e., the personalized recruitment email and the universal link) led participants to a page that provided more information about the study, including data privacy/storage information. By completing the survey, participants provided consent to take part in the study. Participants were not provided financial compensation for taking part in the study. Those who participated in Round 2 of the Delphi were invited to co-author the present paper.

Round 1

In Round 1, participants completed an online Delphi survey that was administered in English on the Welphi platform (*Welphi*, n.d.). The survey was open from 10 July 2024 to 16 August 2024, with reminder emails sent to participants who received personalized email invitations one, two, and four weeks following the initial email. Prior to administration, the survey was pilot tested from 29 June 2024 to 4 July 2024 by four researchers (DS and three external research assistants). Pilot testers did not participate in the Delphi. This pilot test was conducted to check for issues in survey design, technology, and the clarity/phrasing of the survey questions.

The survey included 41 questions that addressed the following: (1) demographic variables (seven close-ended questions); (2) preferences for GLOBAL candidate checklist items (32 questions); and (3) other comments (i.e., suggestions for new items that were not addressed in the GLOBAL and additional comments in general; two open-ended questions). All survey questions were optional to complete, with the exception of rating preferences for the candidate items. Participants were required to complete all the questions on a page to move to the next, but their responses were submitted even if the survey was not fully completed. For the ‘preferences for

GLOBAL candidate items' section, participants were asked to rate each item of the preliminary 32-item GLOBAL checklist that was generated from the scoping review (Ng et al., 2024) using the following Likert scale scoring system (Jebb et al., 2021): essential (1-3), preferable (4-6), and non-essential (7-9). It was determined a priori that items that garnered 80% of responses in the top range (7-9) or bottom range (1-3) on the 9-point scale were considered to have achieved consensus for inclusion or exclusion. This 80% threshold was selected based on general agreement within the literature, which commonly uses 75% as a threshold to define consensus (Diamond et al., 2014). Items that met consensus were excluded from consideration in the subsequent round. Each candidate GLOBAL item in this section also had an open-ended comment box for respondents to provide further feedback. At the end of the survey, participants were provided with information regarding the Round 2 in-person consensus meeting and a linked form to express their interest in attending.

Round 2

A one-day consensus meeting was held on 21 September 2024 in Berlin, Germany, to discuss and vote on new items and those that did not reach consensus in Round 1. The date and location of the meeting were chosen to take advantage of many members of the bibliometric community attending the STI 2024 conference in Berlin from 18 to 20 September 2024. Stakeholders were invited by the steering committee via email from the list of Round 1 Delphi participants who fully completed the survey and expressed interest in participating in Round 2. A total of 32 participants were invited to participate. Efforts were made to ensure varied representation from all stakeholder groups.

The in-person consensus meeting was moderated by three steering committee members (JYN, SH, and LW), who did not vote or participate in discussion but aimed to stay neutral during the meeting. Two researchers (DS and one external research assistant) took notes and recorded votes during this process. During the meeting, all items that did not reach consensus from the initial literature review and all new items proposed by participants in Round 1 were discussed. The consensus group participants were presented with each item along with its score from the first Delphi exercise, in addition to any remarks made by Round 1 participants on that item. This information was provided six days in advance of the meeting on 15 September 2024 as part of a handbook and during the meeting itself on 21 September 2024. At the consensus meeting participants were asked to comment on the significance of each item and whether it should be included in the GLOBAL. After an open discussion of a particular item, participants were given the option to rephrase items if the majority agreed upon its change. After discussions for each given item, an anonymous electronic vote was held using Mentimeter (*Interactive Presentation Software - Mentimeter*, n.d.) with the option to 'include in checklist', 'exclude from checklist', and 'abstain from voting'. After voting for an item was completed, final results were presented quantitatively. Similar to Round 1, the inclusion and exclusion threshold of 80% served to represent majority consensus (Diamond et al., 2014). Participants also had the chance to suggest new items for the GLOBAL during the consensus meeting, and these were subsequently voted on. Participants were not required to

stay for the complete Delphi process given this was a day-long event involving international stakeholders, although this was encouraged. In addition to the notes taken, the meeting was recorded and transcribed using MacWhisper (*MacWhisper*, n.d.).

Analysis

Frequencies and percentages were used to record the number of participants that completed each round of the Delphi and their basic demographic characteristics. For Round 1, qualitative data (open-ended responses) underwent content analysis (Joffe & Yardley, 2003). There were three categories of open-ended responses from the Round 1 survey: 1) item-specific responses (32 questions); 2) suggestions for new items that were not addressed in the GLOBAL (one question); and 3) additional comments in general (one question). Coding to identify common themes in participant responses was conducted independently and in duplicate by two researchers (MM and NS), before meeting to resolve discrepancies in coding. Following this, MM and NS met to iteratively generate and discuss themes and subthemes until consensus was reached. All 'item-specific responses' were reviewed and discussed, but only items deemed to have sufficient data, as determined by team discussion (JYN, HL, MM, NS), were analyzed (e.g., items that had less than three dissimilar comments were determined to have insufficient data). Item-specific responses were coded with the purpose of identifying ways to rephrase items on the GLOBAL for the Round 2 consensus meeting and to capture any concerns regarding GLOBAL items. Responses for 'suggestions for new items that were not addressed in the GLOBAL' were coded with the intention to identify new items to add to the GLOBAL checklist. Newly proposed items were subsequently presented to the research team and steering committee for further refinement. Through iterative team discussions, new items reached consensus for inclusion to vote on during Round 2. Responses from 'additional comments in general' were used to generate general themes regarding participant preferences on the GLOBAL's format and usage and were subsequently presented to participants taking part in the Round 2 Delphi.

Results

The results of each stage of the process of developing the GLOBAL are summarized in Figure 1 and described in more detail in the subsequent sections.

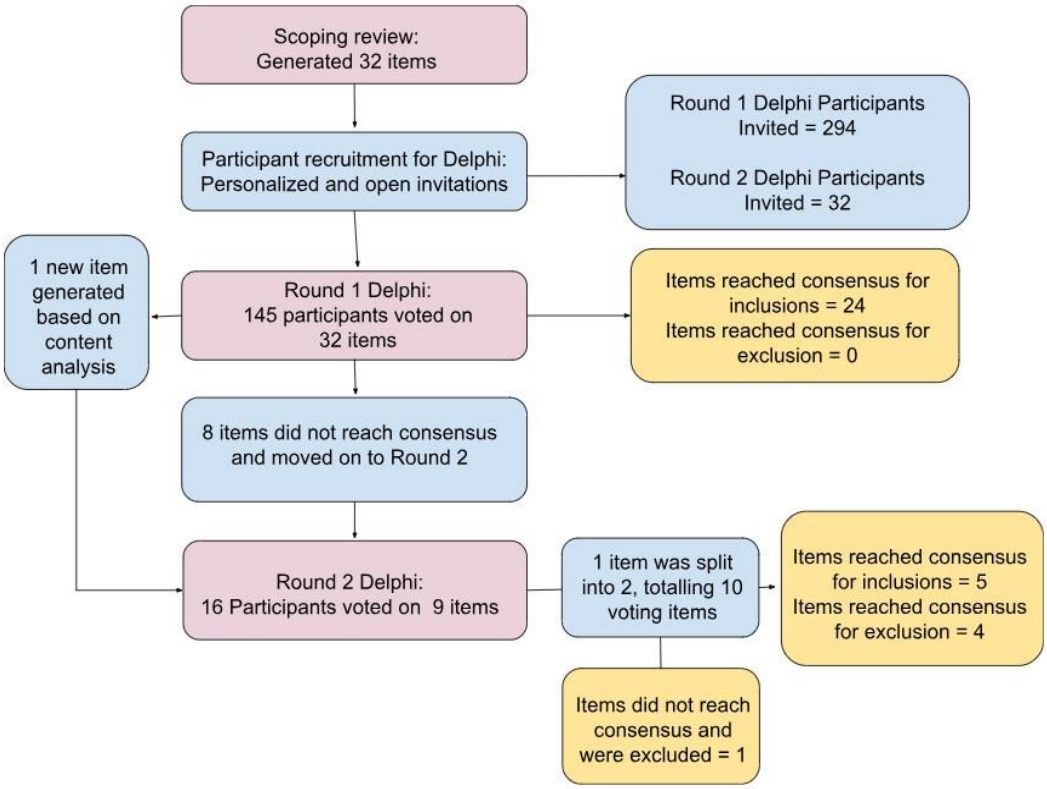


Figure 1. Summary of the methods and results of the GLOBAL development process.

Deviance from the protocol

Time and resource constraints led to three deviations from the protocol. First, the three-round Delphi process was reduced to two rounds as items reached consensus within the two rounds. Second, although Round 2 was initially planned to be an online survey, it was conducted as an in-person consensus meeting instead. While this deviation limited the number of participants who could attend Round 2, it also promoted a productive and detailed discussion for each item. Third, Welphi (*Welphi*, n.d.) was used to create and distribute the Round 1 Delphi survey instead of SurveyLet (*Calibrum*, n.d) since Welphi was designed to implement the Delphi method and ensure data accuracy.

Round 1

Participants

A total of 145 participants, representing 111 institutions, took part in Round 1 by rating at least one GLOBAL item. Table 1 provides a summary of participant demographics. Only two (1.4%) participants did not fully complete the survey. Most respondents were men (n = 91, 62.8%) and between the ages of 35 and 44 (n = 56, 41.4%). Respondents worked in various countries, including the United States (n = 19, 16.0%), Canada (n = 15, 12.6%), Germany (n = 9, 7.6%), the United Kingdom (n = 9, 7.6%), and the Netherlands (n = 8, 6.7%). The top five roles reported by the

participants who completed the survey were: 'bibliometrician' (n = 48, 22.6%), 'librarian/information specialist' (n = 36, 17.0%), 'associate professor' (n = 18, 8.5%), 'full professor' (n = 17, 8.0%), and 'research coordinator' (n = 13, 6.1%). More than half of respondents had more than ten years of experience in their respective careers (n = 70, 57.9%), a quarter had five to ten years of experience (n = 31, 25.6%), and 14.1% had less than five years (n = 17).

Table 1. Round 1 and 2 participant demographic characteristics.

Demographic	*Participant characteristics*	*Responses (n, %)*	
		Round 1	*Round 2*
Gender	Male	91 (62.8%)	11 (68.6%)
	Female	47 (28.9%)	4 (25.0%)
	Prefer not to say	7 (4.8%)	1 (6.2%)
		N=145	N=16
Age	25-29	2 (1.5%)	0 (0.0%)
	30-34	12 (8.9%)	3 (18.8%)
	35-39	28 (20.7%)	2 (12.5%)
	40-44	28 (20.7%)	2 (12.5%)
	45-49	16 (11.8%)	1 (6.2%)
	50-54	15 (11.1%)	2 (12.5%)
	55-59	13 (9.6%)	3 (18.8%)
	60-64	8 (5.9%)	1 (6.2%)
	65-69	6 (4.4%)	2 (12.5%)
	70+	6 (4.4%)	0 (0.0%)
	Prefer not to say	1 (0.7%)	0 (0.0%)
		N=135	N=16
Country of work	Canada	15 (12.6%)	0 (0.0%)
	Germany	9 (7.6%)	4 (28.6%)
	Netherlands	8 (6.7%)	2 (14.3%)
	United Kingdom	9 (7.6%)	0 (0.0%)
	United States	19 (16.0%)	1 (7.1%)
	Other	56 (47.1%)	7 (50.0%)
	Prefer not to say	3 (2.5%)	0 (0.0%)
		N=119	N=14
Career stage	Early (≤5 yrs)	17 (14.1%)	1 (7.1%)
	Mid (5-10 yrs)	31 (25.6%)	3 (21.4%)
	Senior (10+ yrs)	70 (57.9%)	10 (71.4%)
	Prefer not to say	3 (2.5%)	0 (0.0%)
		N=121	N=14
Role	Bibliometrician	48 (22.6%)	7 (26.9%)
	Librarian/ Information specialist	36 (17.0%)	3 (11.5%)
	Research coordinator	13 (6.1%)	2 (7.7%)
	Associate Prof.	18 (8.5%)	3 (11.5%)
	Full Professor	17 (8.0%)	2 (7.7%)
	Other	80 (37.7%)	9 (32.6%)
		N=212[a]	N=26[a]

[a] Participants chose more than one option.

GLOBAL item preferences

A total of 24 out of 32 items reached the 80% consensus threshold for inclusion in the GLOBAL reporting guideline in Round 1. Content analysis of participant feedback resulted in one novel candidate item for inclusion in the GLOBAL: "Provide a clear study materials and data sharing statement (e.g., if datasets, data sources, codes used for the analysis, software, and/or calculations are provided or not)". This item was voted on in Round 2. In the open-ended survey responses, participants suggested two further themes: 1) expanding the GLOBAL objective by adapting it to different audiences and/or types of records that use bibliometric analyses; and 2) reformatting the GLOBAL checklist to reduce redundancy, include examples, or rephrase existing items. A summary of these themes is provided in Table 2. The first theme is discussed and encouraged in the 'Future Directions' section, while the second theme could not be implemented as participants were only allowed to vote on the necessity of the items rather than modify their content.

Table 2. Round 1 Delphi online consensus group themes.

Theme	*Codes*	*Quotes*
GLOBAL objectives	Clarify purpose	"Why is this needed at this time?" (P15) "The meaning of 'reporting bibliometric studies' could be broad. It is unclear whether it specifically apply to reports, [...] research articles, or any document based on a bibliometric analysis." (P77)
	Adapt to audience	"I wonder if the reporting needs to be adapted to the audience" (P156) "I sometimes found it difficult to answer the questions because I produce different types of analysis for very different audiences" (P122)
	Extensions	"How about other types of reporting, e.g. benchmarking reports that institutions and governments use" (P342)
GLOBAL formatting and structure	Redundant / generic requirements	"Some of the items seem to be overlapping in meaning, causing unnecessary redundancy" (P241) "Many of these items seem not particular to bibliometric studies but rather standard elements of journal articles" (P378)
	Include examples	"It would be nice to see some "recipes" (representative examples)." (P320)
	Item editing / suggestions	"[...] the question of data availability and the conflict of interest [...] should be better defined." (P25) "By making all this mandatory one runs the risk of making papers heavy and impenetrable" (P360)

Round 2

Participants

A total of 16 participants took part in Round 2. Demographic characteristics were collected from Round 1, anonymized, and aggregated. Participants were mostly men (n = 11, 68.8%), 'White' (n = 13, 81.3%), and all participants were between 30 and 59 years of age (n = 16, 100%). Most respondents worked in Germany (n = 4, 25.0%),

Denmark (n = 3, 18.8%), and the Netherlands (n = 2, 12.5%). Common roles included 'bibliometrician' (n = 7, 25%), 'journal editor' (n = 3, 10.7%), 'librarian/information specialist' (n = 3, 10.7%), and 'associate professor' (n = 3, 10.7%). Most participants had more than ten years (n = 10, 62.5%), and few with between five and ten (n = 3, 18.8%) and less than five years (n = 1, 6.4%) of work experience. Summarized participant demographics are provided in Table 1.

GLOBAL item preferences

Participants voted on ten items during the in-person consensus meeting. Initially, there were nine items from Round 1 that required further discussion, eight of which did not reach consensus and one that was introduced after content analysis of participant feedback. However, one item was split into two during Round 2, resulting in ten items. The phrasing of seven items was altered. Five out of ten items reached the 80% consensus threshold for inclusion following Round 2, with all five undergoing rephrasing. The other five items were excluded, with four of them reaching consensus for exclusion. The remaining item did not reach the consensus threshold and was therefore excluded. Participants did not suggest any new items for the GLOBAL. In total, 29 out of 34 items reached consensus for inclusion in the GLOBAL following the completion of both Delphi rounds. A summary of the original items included and their results via the two Delphi rounds is provided in the Appendix.

Guidelines finalization process

The final GLOBAL checklist is comprised of 29 items, with the following in each section: 'abstract' (one item), 'introduction' (four items), 'methods' (13 items), 'results' (four items), 'discussion' (three items), and 'other' (four items). The finalized GLOBAL checklist is presented in Table 2.

Table 2. The final 29-items GLOBAL guideline for the reporting of bibliometric analyses.

	Reporting item
Abstract	
1.1	Abstract should be reflective of the bibliometric analysis, including scope, data collection, analysis, and results.
Introduction	
2.1	Situate the bibliometric analysis within the context of relevant pre-existing literature, identifying the gap in literature.
2.2	Define the aim, scope, rationale, and/or objective of the bibliometric analysis.
2.3	Define the research question.
2.4	Explicitly specify relevant terms, concepts, and theoretical frameworks used in the study.
Methods	
3.1	Describe the bibliometric methods used.
3.2	Define the units of analysis that are analysed (i.e., micro-, meso-, and macro-level) in the bibliometric analysis (e.g., countries, institutions, authors).

3.3	Describe the bibliometric data collection methods, including any limitations.
3.4	Describe the databases and data sources used, including any limitations.
3.5	Present the full search strategies for all databases used, including any filters and limits that were applied.
3.6	Describe the data collection time frame.
3.7	Describe the search results and selection processes (e.g., inclusion/exclusion). If applicable, use a flow diagram.
3.8	Describe the data cleaning methods, including any limitations.
3.9	Describe the bibliometric data analysis methods used.
3.10	Specify the analytical software used and the parameter settings selected.
3.11	Describe the bibliometric indicators used.
3.12	If applicable, define the calculations/formulas used for indicators in the bibliometric analysis.
3.13	Provide sufficient detail in the bibliometric analysis manuscript to ensure full replicability/transparency of methods.
Results	
4.1	Describe the results and key findings.
4.2	Describe the results of bibliometric analysis techniques used.
4.3	Ensure figures, tables and visualizations clarify and/or facilitate the interpretation of the results without misleading.
4.4	If appropriate, report the uncertainty/dispersion/heterogeneity depending on the type of data and analysis, and error values of bibliometric indicators.
Discussion	
5.1	Summarize and discuss study findings.
5.2	Provide context for and situate the study findings in the literature.
5.3	Discuss the strengths, limitations, and potential biases of the bibliometric analysis.
Other	
6.1	Disclose any existing or potential conflicts of interest and/or sources of financial or non-financial support.
6.2	Describe the availability and accessibility of data.
6.3	Use references and citations to support statements and methods used.
6.4	Provide a statement about whether study materials, data and/or code are shared and if so, where and how it can be accessed.

Discussion

The GLOBAL serves as the first guideline developed for the reporting of bibliometric analyses in the scholarly literature through international multi-stakeholder and multi-sector consensus. Through an iterative, multi-step process, we have developed a 29-item reporting guideline that is intended to enable more thorough, accurate, and transparent reporting of bibliometric analyses. It is important to note that these are minimum standards. Authors should not feel discouraged from including additional information that might enhance the quality of reporting of their bibliometric analysis.

Scope of GLOBAL

The goal of the GLOBAL is to provide the minimum essential guidance for the reporting of bibliometric analyses for research purposes. The intent of the GLOBAL is not to provide methodological design guidance for researchers and specialists conducting bibliometric analyses, nor does it assess the suitability of particular methods in specific contexts. However, while our work does not directly address the quality of bibliometric analyses, we anticipate that this reporting guideline will set the stage for future work in this area. The complete reporting of novel or more specialized types of bibliometric analyses may require additional guideline items and authors should not be deterred from reporting this information. The GLOBAL should nevertheless be considered as base guideline by such studies, until necessary specialized extensions are developed. The latter may also address the reporting of other "metrics" associated with bibliometrics (e.g., the reporting of altmetrics or other topics nestled within bibliometrics).

The GLOBAL is formatted to support the reporting process of manuscripts intended to be submitted to scholarly journals or preprint servers, and for peer-review. It incorporates the conventional sections of 'abstract', 'introduction', 'methods', 'results', and 'discussion,' along with an 'other' section, within its design. We aimed to ensure that this reporting guideline is clear and easy to follow, as recommended by the Consolidated Standards of Reporting Trials (CONSORT) (Altman, 1996): "[r]eaders should not have to infer what was probably done; they should be told explicitly." Although the GLOBAL aims to ensure the complete reproducibility of bibliometric analyses, we acknowledge that practical considerations (e.g., journal requirements or concision) may prevent researchers from providing the full scope of information needed to meet the ideal standards for reporting.

Implementation and dissemination

The GLOBAL is currently undergoing pilot testing with experts in the bibliometric community to assess the clarity of items' wording and any issues of redundancy or duplication of items when using the guidelines. Further, an Explanation and Elaboration (E&E) document of the GLOBAL is currently under development. The E&E document will facilitate use of the GLOBAL by providing concrete examples from the published bibliometric literature of suitable reporting, and additional information explaining the item and the rationale for its inclusion in the GLOBAL. Once the pilot testing and E&E document are completed, we plan on disseminating our publication(s) to multiple sources, including but not limited to the following: 1) the core bibliometrics and reporting guidelines communities via conferences and/or mailing lists associated with ISSI and STI, the International Network of Research Management Societies, the Directorate for Science, Technology and Innovation, the European Network of Indicator Designers, and the International Congress on Peer Review and Scientific Publication; 2) editors and editorial board members of scholarly journals; 3) researchers from disciplines that use bibliometrics and/or reporting guidelines to evaluate their own fields; 4) scholarly communication librarians and research managers that conduct bibliometric analyses to support researchers; 5) publishers and publishing-related organizations/associations that

publish bibliometric analyses or bibliometric-related studies; 6) websites and blogs that feature bibliometric-related content, such as The Scholarly Kitchen and Leiden Madtrics; 7) developers of applications and software that assist researchers unfamiliar with bibliometric analyses in conducting and reporting them, such as Bibliometrix (*Bibliometrix*, n.d.); and summer schools offering bibliometric-related programs, such as European Summer School for Scientometrics (ESSS) (european summer school for scientometrics, n.d.) and Centre for Science and Technology Studies (CWTS) (CWTS, n.d.).

Future directions

Future studies may include GLOBAL extensions that address the reporting of other "metrics" associated with bibliometrics, such as webometrics and altmetrics. Additionally, while the current focus is on reporting in the scholarly literature, such as in journal articles, it would also be valuable to develop reporting guidelines for other types of bibliometric analyses, such as analyses performed for research institutions, research funders, governments, and other stakeholders, for instance in a research assessment context. Thus, in the future, extensions of the GLOBAL could be developed that would support authors in reporting bibliometric analyses for the purposes of policy reports, institutional benchmarking, funding evaluations, and other applications. Future research could also explore the development of reporting guidelines for studies that use bibliometrics along with other methodological approaches, such as systematic reviews.

Further research may also examine the facilitators and barriers to the use of the GLOBAL by authors, editors, and peer reviewers, and develop interventions to overcome identified barriers and evaluate those interventions. Moreover, conducting think-aloud studies to understand how items are interpreted and reliability studies to identify where items can be differently interpreted would be beneficial to inform potential revisions to the guideline (Charters, 2003).

Multiple translations of the reporting guideline will improve the accessibility of the GLOBAL. We encourage journal editors and publishers to promote the GLOBAL (for instance, by mentioning it in their journal's "Instructions to Authors" page), endorse its usage, advise editors and peer reviewers to assess submitted bibliometric analyses against the GLOBAL, and adjust journal policies to take into account the new reporting recommendations.

Strengths and limitations

This study has several strengths. First, the development of the GLOBAL adheres to recommendations present within the EQUATOR toolkit and other established guidelines for developing a reporting guideline (EQUATOR Network, n.d.; Moher et al., 2010), thereby increasing its robustness. Second, the development process is evidence-based, supported by a comprehensive scoping review (Ng et al., 2024) of recommendations in the literature. Third, involving diverse stakeholders from the international community (e.g., researchers with varying years of experience with bibliometrics) in the selection process strengthens the study's credibility and relevance as it considers a wide range of perspectives. Fourth, the recruitment of

participants through two methods, sending personalized emails and issuing an open invitation through public advertisement, helped to minimize the potential for bias that could arise from selecting individual participants or relying on a single sampling method.

Conversely, weaknesses of the study include a possible decrease in representativeness due to English-language restrictions, which limited participation by non-English speakers (Khanna et al., 2022). The in-person consensus meeting in Berlin is another limitation, as not all stakeholders were able to attend the meeting and provide feedback on the GLOBAL checklist items, thereby potentially restricting participant diversity. In future, for instance during the development of extensions of the GLOBAL, such meetings could be held in hybrid or virtual formats to facilitate broader participation.

Conclusions

The GLOBAL serves as a guide for high-quality reporting of bibliometric analysis. We anticipate that the GLOBAL checklist will be useful to bibliometricians, librarians, policy and research analysts, and researchers, as well as authors, editors, and peer reviewers of bibliometric analyses. Ultimately, the goal of the GLOBAL is to promote more thorough, accurate, and transparent reporting of bibliometric analyses.

Acknowledgements

We gratefully acknowledge Anton Ninkov, Chantal Ripp, and Marc-Andre Simard for pilot-testing the Welphi survey for Round 1 of the Delphi process. We greatly appreciate and acknowledge the efforts of the Round 1 participants who completed the survey. We also gratefully acknowledge Chantal Ripp for taking notes during Round 2 of the Delphi process. We gratefully acknowledge Stephan Gauch for providing technical equipment and organizing the meeting space and recording and transcription of Round 2. We further gratefully acknowledge Dalia Zubashev for her assistance in editing this submission.

Competing interests

The authors declare that they have no competing interests.

Funding

JYN's postdoctoral fellowship was funded by a MITACS Elevate Award (Award #: IT36020), and co-funded by EBSCO Health. We also gratefully acknowledge funding provided by Cabells. Additionally, we thankfully acknowledge the Korean Institute of Oriental Medicine for their support. The funders played no role in the study design and conceptualization, data collection and analysis, decision to publish, or preparation of this research.

Research ethics approval and transparency practices

Ethics approval was obtained by the Ottawa Health Science Network Research Ethics Board (REB ID #20230527-01H).

References

AlRyalat, S. A. S., Malkawi, L. W., & Momani, S. M. (2019). Comparing Bibliometric Analysis Using PubMed, Scopus, and Web of Science Databases. Journal of Visualized Experiments (JoVE), 123, e55617. https://doi.org/10.3791/58494

Altman, D. G. (1996). Better reporting of randomised controlled trials: The CONSORT statement. BMJ, 313(7057), 570–571. https://doi.org/10.1136/bmj.313.7057.570

Bibliometrix.(n.d.). Retrieved November 17, 2024, from https://www.bibliometrix.org/home/

Bornmann, L., Guns, R., Thelwall, M., & Wolfram, D. (2021). Which aspects of the Open Science agenda are most relevant to scientometric research and publishing? An opinion paper. Quantitative Science Studies, 2(2), 438–453. https://doi.org/10.1162/qss_e_00121

Calibrum. (n.d.). Retrieved September 28, 2024, from https://calibrum.com/

Charters, E. (2003). The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. Brock Education Journal, 12(2), Article 2. https://doi.org/10.26522/brocked.v12i2.38

CWTS: CWTS Scientometrics Summer School. (n.d.). CWTS. Retrieved January 24, 2025, from https://www.cwts.nl/education/cwts-course-program/cwts-scientometrics-summer-school/

De Bellis, N. (2014). 2: History and Evolution of (Biblio)Metrics. Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. https://doi.org/10.7551/mitpress/9445.003.0004

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. Journal of Clinical Epidemiology, 67(4), 401–409. https://doi.org/10.1016/j.jclinepi.2013.12.002

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. Journal of Business Research, 133, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

EQUATOR Network. (n.d.). Retrieved September 28, 2024, from https://www.equator-network.org/

EQUATOR Network. (n.d.a). Retrieved February 5, 2025 from https://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-other-study-designs/.

European summer school for scientometrics – esss.info. (n.d.). Retrieved January 24, 2025, from https://esss.info/

Gagnier, J. J., Kienle, G., Altman, D. G., Moher, D., Sox, H., & Riley, D. (2013). The CARE Guidelines: Consensus-based Clinical Case Reporting Guideline Development. Global Advances in Health and Medicine, 2(5), 38–43. https://doi.org/10.7453/gahmj.2013.008

Gattrell, W. T., Logullo, P., Van Zuuren, E. J., Price, A., Hughes, E. L., Blazey, P., et al. (2024). ACCORD (ACcurate COnsensus Reporting Document): A reporting guideline for consensus methods in biomedicine developed via a modified Delphi. PLOS Medicine, 21(1), e1004326. https://doi.org/10.1371/journal.pmed.1004326

GLOBAL Delphi Survey. (n.d.). Retrieved November 30, 2024, from https://sti2024.org/sti-conference/global-delphi-survey/

Interactive presentation software—Mentimeter. (n.d.). Retrieved September 28, 2024, from https://www.mentimeter.com/

Jappe, A. (2020). Professional standards in bibliometric research evaluation? A meta-evaluation of European assessment practice 2005–2019. PLOS ONE, 15(4), e0231735. https://doi.org/10.1371/journal.pone.0231735

Jebb, A., Ng, V., & Tay, L. (2021). A Review of Key Likert Scale Development Advances: 1995–2019. Frontiers Psychology, 12. https://doi.org/10.3389/fpsyg.2021.637547

Joffe, H., & Yardley, L. (2003). Research Methods for Clinical and Health Psychology. Content and Thematic Analysis, 56–68.

Khanna, S., Ball, J., Alperin, J.P., &Willinsky, J. (2022). Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. Quantitative Science Studies, 3(4), 912–930. doi: https://doi.org/10.1162/qss_a_00228

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. Australian Journal of Management, 45(2). https://doi.org/10.1177/0312896219877678

MacWhisper. (n.d.). Gumroad. Retrieved September 28, 2024, from https://goodsnooze.gumroad.com/l/macwhisper

Moher, D. (2007). Reporting research results: A moral obligation for all researchers. Canadian Journal of Anesthesia, 54(5), 331–335. https://doi.org/10.1007/BF03022653

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010). Guidance for Developers of Health Research Reporting Guidelines. PLOS Medicine, 7(2), e1000217. https://doi.org/10.1371/journal.pmed.1000217

Ng, J. Y., Haustein, S., Ebrahimzadeh, S., Chen, C., Sabé, M., Solmi, M., & Moher, D. (2023). Guidance List for repOrting Bibliometric AnaLyses (GLOBAL). https://doi.org/10.17605/OSF.IO/MTXBF

Ng, J. Y., Liu, H., & Haustein, S. (2024). The GLOBAL initiative - Contribute to improving the reporting of bibliometric analyses. Leiden Madtrics. Retrieved September 28, 2024, from https://www.leidenmadtrics.nl/articles/the-global-initiative-contribute-to-improving-the-reporting-of-bibliometric-analyses

Ng, J. Y., Liu, H., Masood, M., Syed, N., Stephen, D., Ayala, A. P., Sabé, M., Solmi, M., Waltman, L., Haustein, S., & Moher, D. (2024). Guidance for the Reporting of Bibliometric Analyses: A Scoping Review (p. 2024.08.26.24312538). medRxiv. https://doi.org/10.1101/2024.08.26.24312538

Peters, M. D. J., Godfrey, C., McInerney, P., et al. (2020). Chapter 11: Scoping Reviews (2020 version). In E. Aromataris & Z. Munn (Eds.), JBI Manual for Evidence Synthesis (pp. 406–451). JBI. https://doi.org/10.46658/JBIMES-20-01

Resnik, David. B. (2015). What Is Ethics in Research and Why Is It Important? National Institute of Environmental Health Sciences. https://www.niehs.nih.gov/research/resources/bioethics/whatis

Sugimoto, C. R., & Larivière, V. (2018). Measuring Research: What Everyone Needs to Know®. Oxford University Press.

Welphi. (n.d.). Retrieved September 28, 2024, from https://www.welphi.com/en/Home.htmluckland, M. & Gey, F. (1994). The relationship between recall and precision. *Journal of the American Society for Information Science*, 45, 12-19.

Appendix

Summary of Delphi voting rounds

Section	Checklist item [a]			Agreement [b] (n[%])		Final outcome
	Preliminary	Round 1	Round 2	Round 1[c]	Round 2[c]	
Title	In the title, identify the study as a bibliometric analysis and indicate the time period and key issues/topic.	In the title, identify the study as a bibliometric analysis and indicate the time period and key issues/topic.	In the title, identify the study as a bibliometric analysis and indicate the time period and key issues/topic.	Essential (1-3): 74 (51.03%) Neutral (4-6): 57 (39.31%) Non-Essential (7-9): 14 (9.66%)	Include: 3 (18.75%) **Exclude: 13 (81.25%)** Abstain: 0 (0%)	Excluded
Abstract	Abstract should be reflective of the bibliometric analysis, including scope, data collection, analysis, and results.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 120 (83.33%)** Neutral (4-6): 23 (15.97%) Non-Essential (7-9): 1 (0.69%)	Consensus reached in Round 1	Included
Introduction	Situate the bibliometric analysis within the context of relevant pre-existing literature, identifying the gap in literature.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 117 (81.25%)** Neutral (4-6): 22 (15.28%) Non-Essential (7-9): 5 (3.47%)	Consensus reached in Round 1	Included
Introduction	Define the aim, scope, rationale, and/or objective of the bibliometric analysis.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 137 (95.80%)** Neutral (4-6): 6 (4.20%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included

Introduction	Define the research question.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 131 (91.61%)** Neutral (4-6): 12 (8.39%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Introduction	Clearly define all relevant terms and definitions used within the bibliometric analysis.	Clearly define all relevant terms and definitions used within the bibliometric analysis.	Explicitly specify relevant terms, concepts, and theoretical frameworks used in the study.	Essential (1-3): 104 (72.73%) Neutral (4-6): 35 (24.48%) Non-Essential (7-9): 4 (2.80%)	**Include: 16 (100%)** Exclude: 13 (81.25%) Abstain: 0 (0%)	Included
Introduction	Describe the intended target audience of the bibliometric analysis (e.g., researchers, public, media, etc.). Describe the ways in which the information included in the report may be used for the target audience.	Describe the intended target audience of the bibliometric analysis (e.g., researchers, public, media, etc.). Describe the ways in which the information included in the report may be used for the target audience.	[Rephrased into two items] Item #1: Describe the intended target audience of the bibliometric analysis (e.g. researchers, public, media, etc).	Essential (1-3): 56 (39.16%) Neutral (4-6): 76 (53.15%) Non-Essential (7-9): 11 (7.63%)	Include: 1 (6.25%) **Exclude: 15 (93.75%)** Abstain: 0 (0%)	Excluded
			[Rephrased into two items] Item #2: Describe the ways in which the information included in the report is expected to be of relevance or intended to be used.		Include: 3 (6.25%) **Exclude: 10 (93.75%)** Abstain: 3 (0%)	
Methods	Describe the bibliometric methods used.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 137 (95.74%)** Neutral (4-6): 6 (4.26%)	Consensus reached in Round 1	Included

				Non-Essential (7-9): 0 (0.00%)		
Methods	Define the units of analysis that are analysed (i.e., micro-, meso-, and macro-level) in the bibliometric analysis (e.g., countries, institutions, authors).	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 126 (90.00%)** Neutral (4-6): 14 (10.00%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Methods	Describe the bibliometric data collection methods, including any limitations.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 137 (97.86%)** Neutral (4-6): 3 (2.14%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Methods	Describe the databases and data sources used, including any limitations.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 137 (97.86%)** Neutral (4-6): 2 (1.43%) Non-Essential (7-9): 1 (0.71%)	Consensus reached in Round 1	Included
Methods	Present the full search strategies for all databases used, including any filters and limits that were applied.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 124 (88.57%)** Neutral (4-6): 16 (11.43%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Methods	Describe the data collection time frame.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 126 (90.00%)** Neutral (4-6):	Consensus reached in Round 1	Included

				14 (10.00%) Non-Essential (7-9): 0 (0.00%)		
Methods	Describe the search results and selection processes (e.g., inclusion/exclusion). If applicable, use a flow diagram.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 114 (81.43%)** Neutral (4-6): 26 (18.57%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Methods	Describe the data cleaning methods, including any limitations.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 120 (85.71%)** Neutral (4-6): 19 (13.57%) Non-Essential (7-9): 1 (0.71%)	Consensus reached in Round 1	Included
Methods	Describe the bibliometric data analysis methods used.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 134 (95.71%)** Neutral (4-6): 6 (4.29%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Methods	Specify the analytical software used and the parameter settings selected.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 116 (82.86%)** Neutral (4-6): 24 (17.14%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included

Methods	Describe the bibliometric indicators used.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 128 (91.43%)** Neutral (4-6): 12 (8.57%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Methods	If applicable, define the calculations/formulas used for indicators in the bibliometric analysis.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 115 (82.14%)** Neutral (4-6): 25 (17.86%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Methods	Provide sufficient detail in the bibliometric analysis manuscript to ensure full replicability / transparency of methods.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 118 (84.29%)** Neutral (4-6): 22 (15.71%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Results	Describe the results and key findings.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 136 (97.14%)** Neutral (4-6): 4 (2.86%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Results	Describe the results of bibliometric analysis techniques used.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 123 (87.86%)** Neutral (4-6): 15 (10.71%) Non-Essential (7-9):	Consensus reached in Round 1	Included

				2 (1.43%)		
Results	Visualize the results through the use of figures, graphs, and/or tables. Ensure the visualizations are simple and easy to interpret. Aesthetic bibliometric visualization should not replace a rigorous bibliometric analysis.	Visualize the results through the use of figures, graphs, and/or tables. Ensure the visualizations are simple and easy to interpret. Aesthetic bibliometric visualization should not replace a rigorous bibliometric analysis.	Ensure figures, tables and visualizations clarify and/or facilitate the interpretation of the results without misleading.	Essential (1-3): 102 (72.86%) Neutral (4-6): 33 (23.57%) Non-Essential (7-9): 5 (3.57%)	**Include: 13 (86.67%)** Exclude: 0 (0%) Abstain: 2 (13.33%)	Included
Results	If applicable, report the uncertainty / dispersion/heterogeneity depending on the type of analysis and error values of bibliometric indicators.	If applicable, report the uncertainty /dispersion/heterogeneity depending on the type of analysis and error values of bibliometric indicators.	If appropriate, report the uncertainty/ dispersion/heterogeneity depending on the type of data and analysis, and error values of bibliometric indicators	Essential (1-3): 97 (69.29%) Neutral (4-6): 43 (30.71%) Non-Essential (7-9): 0 (0.00%)	**Include: 12 (80%)** Exclude: 2 (13.33%) Abstain: 1 (6.67%)	Included
Discussion	Summarize and discuss study findings.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 129 (92.14%)** Neutral (4-6): 11 (7.86%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Discussion	Elaborate on the applicability and implications of study findings.	Elaborate on the applicability and implications of study findings.	Discuss the applicability and implications of study findings.	Essential (1-3): 105 (75.00%) Neutral (4-6): 34 (24.29%) Non-Essential (7-9): 1 (0.71%)	Include: 9 (60%) Exclude: 5 (33.33%) Abstain: 1 (6.67%)	Excluded[d]

Discussion	Provide context for the results of the bibliometric analysis and situate the study findings in existing literature.	Provide context for the results of the bibliometric analysis and situate the study findings in the existing literature.	Provide context for and situate the study findings in the literature.	Essential (1-3): 104 (74.29%) Neutral (4-6): 35 (25.00%) Non-Essential (7-9): 1 (0.71%)	**Include: 13 (86.67%)** Exclude: 2 (13.33%) Abstain: 0 (0%)	Included
Discussion	Discuss the strengths, limitations, and potential biases of the bibliometric analysis.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 128 (90.00%)** Neutral (4-6): 13 (9.29%) Non-Essential (7-9): 1 (0.71%)	Consensus reached in Round 1	Included
Discussion	Identify future directions for research.	Identify future directions for research.	Identify future directions for research.	Essential (1-3): 54 (38.57%) Neutral (4-6): 79 (56.43%) Non-Essential (7-9): 7 (5.00%)	Include: 3 (20%) **Exclude: 12 (80%)** Abstain: 0 (0%)	Excluded
Other	Disclose any existing or potential conflicts of interest and/or sources of financial or non-financial support.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 112 (80.00%)** Neutral (4-6): 27 (19.29%) Non-Essential (7-9): 1 (0.71%)	Consensus reached in Round 1	Included
Other	Describe the availability and accessibility of data.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 114 (81.43%)** Neutral (4-6): 26 (18.57%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included

Other	Use references and citations to support statements and methods used.	Consensus reached in Round 1	Consensus reached in Round 1	**Essential (1-3): 125 (89.29%)** Neutral (4-6): 15 (10.71%) Non-Essential (7-9): 0 (0.00%)	Consensus reached in Round 1	Included
Other	[Not Included in Round 1]	Provide a clear study materials and data sharing statements (e.g. if datasets, data sources, codes used for the analysis, software, and/or calculations are provided or not).	Provide a statement about whether study materials, data and/or code are shared and if so, where and how it can be accessed.	[Not Included in Round 1]	**Include: 14 (100%)** Exclude: 0 (0%) Abstain: 0 (0%)	Included

a Underlining denotes text changes made between rounds.

b Bold indicates consensus.

c Round 1 items were scored on a 9-point Likert scale, where 1 to 3 points were categorized as 'essential',' 4 to 7 points were categorized as 'neutral,' and 7 to 9 points were categorized as 'non-essential' for inclusion within the tool. Round 2 items were scored using 'include in checklist', 'exclude from checklist', and 'abstain from voting' for inclusion within the tool.

d Item excluded because 80% threshold for consensus was not reached.

https://doi.org/10.51408/issi2025_023

Higher Standards and Unnoticed Preference - the Impact of Editor-in-Chief on Collaborators

Hao Yueru[1], Yue Mingliang[2], Ma Tingcan[3]

[1]*haoyueru24@mails.ucas.ac.cn*, [2]*yueml@whlib.ac.cn*, [3]*matc@whlib.ac.cn*
National Science Library (Wuhan), Chinese Academy of Sciences, Wuhan 430071 (China)
Department of Information Resources Management, University of Chinese Academy of Sciences, Beijing 100190 (China)

Abstract

Understanding the influence of Editors-in-Chiefs (EiCs) on their collaborators provides valuable insights into the complex interplay between editorial leadership and academic collaboration, shedding light on how such dynamics shape publication practices and journal quality. This study investigated the influence of EiCs' appointment on their collaborators' publishing behaviors in computer science journals listed on ScienceDirect. By employing the Wilcoxon Signed Rank test and T-tests, we analyzed submission Willingness, Share, and Academic Value (of published papers) across three author categories, i.e., Listed Authors, Core Authors, and Other Authors. Results revealed a stable submission willingness but a decline in publication share for Listed and Core Authors post-appointment. Trends in the academic value of articles were mixed: Core Authors showed improvement under stricter standards, while Other Authors experienced a decline (statistically significant at the 90% confidence level). These findings highlight the EiCs' role in balancing editorial rigor and collaborative dynamics, but further research across disciplines is needed due to sample size and research field limitations.

Introduction

Scientific journals serve as critical platforms for scholars to engage in academic exchanges and disseminate their findings, and they gather the original and innovative contributions of science and have a profound social and academic impact (Mauleón et al., 2013).

Editorial Board Members (EBMs) are generally regarded as distinguished researchers with exceptional publication and citation records (Schubert, 2017). As the "gatekeepers of science" (Mauleón et al., 2013; Helgesson et al., 2022; Scarlato et al., 2024), EBMs play a pivotal role in shaping the journal's academic quality. Their primary responsibilities include assessing manuscripts for suitability for the journal (Hames, 2001) and selecting papers with excellent scientific content (Tokić, B. 2017). Moreover, the impact of editorial bias on authors' satisfaction and motivation can influence the types of manuscripts submitted to journals (García et al., 2015).

However, EBMs are not only gatekeepers but also contributors to the research ecosystem, often participating as authors and collaborators themselves. This dual role can lead to potential conflicts of interest, including perceived or actual biases involving close collaborators, research partners, or co-authors (ICMJE, 2024; COPE, 2024; CSE, 2024). "Publication bias" remains a broadly perceived preconception (Mani et al., 2013). To address these issues, several studies had explored the

influence of EBMs' co-authorship on journal outcomes (Colussi, 2018; Ductor & Visser, 2022).

Considering that Editors-in-Chief (EiCs) are the top decision-makers of journals, many scholars have embarked on an exploration of the "self-publishing" phenomenon associated with EiCs (Liu et al., 2023; Nourmand et al., 2024). It has been observed that some EiCs have self-publishing rates that are relatively elevated in comparison to those of other editors (Liu et al., 2023). Additionally, within the context of several dental journals, a substantially increased number of self-publications has been detected, which consequently engenders potential conflicts of interest for EiCs (Nourmand et al., 2024). Meanwhile, the potential conflicts of interest arising from the collaborative relationships of EiCs have yet to be fully explored.

In this paper, aiming to explore the impact of the EiCs on collaborators, we selected collaborators based on the frequency of previous co-authorships, classified the author types in the article, and analyzed changes in their publication willingness, share, and academic value before and after the EiCs' appointment, to investigate whether there are potential conflicts of interest between the EiCs and their collaborators before and after the EiCs' appointment.

The remainder of this paper is as follows. *Related work* introduces previous research related to our study. *Data and methods* describes the process of dataset construction and the definition of observation indicators. *Results* shows the results of analysis and *Discussion* give some discussions.

Related work

The Editor-in-Chief (EiC), or an equivalent with a similar title, is the top decision-maker in academic journals (Schubert, 2017), and holds substantial influence over the journal's editorial policies, submission practices, and overall quality. EiCs are responsible for both maintaining high standards of excellence and overseeing journal operations (Nourmand et al., 2024), as well as improving the quality and impact of the journals they edit. Previous studies can be primarily divided into two types, one incorporates EiCs into the scope of editorial board members (EBMs) for research purposes, the other conducts research on EiCs as a distinct cohort.

The phenomenon of self-publishing by EiCs has been extensively studied, revealing its contentious nature and significant variation across disciplines. Helgesson et al. (2022) highlighted the heterogeneity of editorial influence reflected in differing self-publishing rates among journals. Zdeněk (2018) found that the share of articles authored by editorial board members (EBMs) in their own journals is positively correlated with the gap between impact factor and impact factor without Journal Self Cites, and negatively correlated with the Article Influence Score. Similarly, Zdeněk and Lososová (2018) observed that in agricultural economics journals, higher self-publishing rates among EBMs inversely correlated with bibliometric indicators such as uncited articles.

In contrast, Walters (2015) reported that 64% of EBMs in library and information science journals published fewer articles than expected, potentially reflecting efforts to avoid conflicts of interest. Scanff et al. (2021) identified editorial bias through

analysis of prolific authors and Gini indices in biomedical journals, where 26% of the most prolific authors were EiCs. Liu et al. (2023) examined 81,000 editors over five decades across 15 disciplines, finding that EiCs tend to self-publish at higher rates. Furthermore, Nourmand et al. (2024) quantified self-publications in dental journals and reported a significant increase in potential conflicts of interest. These studies collectively underscore the complex dynamics and implications of self-publishing by EiCs.

Beyond their own publishing habits, EiCs also influence the publication outcomes of collaborators. Research has demonstrated how personal and professional connections between authors and EBMs can influence publication decisions. Colussi (2018) explored how different types of connections—such as shared faculty membership, common PhD advisors, or co-authorship history—affect the quality of published papers. And the findings suggest that connections ultimately improve the quality of published papers, the share of Co-authors connected papers is around 8%. In the view of co-authors, there is no obvious increase in their publication outcomes when this editor is in charge of a journal. Ductor & Visser (2022) investigated the situation when a coauthor joins an editorial board. They found when the coauthor joins an editorial board of an economics journal, the scholar publishes more articles in the coauthor's journal, and point that more editorial power over submissions means larger increases.

Further study by Sarigöl et al. (2017) showed that prior co-authorship with an editor can significantly reduce manuscript handling times, demonstrating that personal relationships can expedite the editorial decision-making process. Trieschmann et al. (2000) and Brogaard et al. (2014) also showed that faculty members at universities with faculty serving as editors tend to have increased publication output. Trieschmann et al. (2000) found that business schools with faculty holding editorial positions in journals saw improved research performance, while Brogaard et al. (2014) observed that faculty at the editor's university published twice as many papers during the editor's tenure compared to when the faculty member was not serving as editor.

The impact of EBMs' personal relationships with authors has been the subject of some discussion. Some scholars have contended that such practices may improve the efficiency of the academic publishing process. Laband and Piette (1994) suggested that what many consider "favoritism" might actually serve to enhance efficiency in the market for scientific knowledge. By favoring collaborations with established researchers, editors may streamline the editorial process and improve the quality of publications. Colussi (2018) also found that the social connections ultimately improve the quality of published papers. Therefore, while personal relationships in editorial decisions may appear biased, they can also contribute to better journal quality and greater research dissemination.

The existing body of research emphasizes the multifaceted influence of EiCs on academic publishing, particularly concerning self-publishing practices, editorial bias, and their impact on collaborators. These studies offer valuable insights into the editorial dynamics and underscore the dual role of EiCs as gatekeepers and contributors to the research ecosystem. However, while prior research has primarily

focused on the prevalence of self-publishing and general trends in editorial influence, the nuanced effects of an EiC's appointment on collaborators' publishing behaviors, including their willingness to submit, publication share, and the academic value of their papers, have not been sufficiently studied. This study aims to investigate these aspects through a focused analysis of computer science journals, employing rigorous statistical methods to reveal trends across different collaborator roles. By connecting these insights with broader editorial practices, this research not only complements the existing literature but also offers a novel perspective on the balance between editorial rigor and collaborative dynamics under the EiCs leadership.

Data and methods

This study aims to analyze the potential changes in the collaborators' publishing practices before and after the appointment of the corresponding EiCs. For this purpose, the bibliographic data of EiCs and their collaborates was collected and analyzed. Fig. 1 gives the framework of the work, which includes data collection, variable definition and statistical analysis.

Data collection

The first thing is to determine the research object, i.e., EiCs and collaborators, based on which the bibliographic data can be collected. The determination of the EiCs is subject to 2 criteria. First, data accessibility. We need to collect the EiCs' names, affiliations and appointment periods, which are crucial for subsequent analysis. After reviewing various journal platforms, we finally chose the ScienceDirect database for its extensive and openly accessible editorial board information. Typically, editorial board details, including the EiCs' name, affiliation and position, can be found in the front matter of journal issues. ScienceDirect provides the information for most journals as free-access PDF files, which can be easily downloaded for the analysis. Second, time restrictions. According to Colussi (2018), a six-year window is well-suited for observing the bibliometric changes related to the appointment of EiCs. Our analysis also used the six-year window, three years before and three years after the EiCs' appointment, to examine the potential changes. That makes the appointment year of an EiC should not be later than 2019 (the initial data collection time is 2024.6).

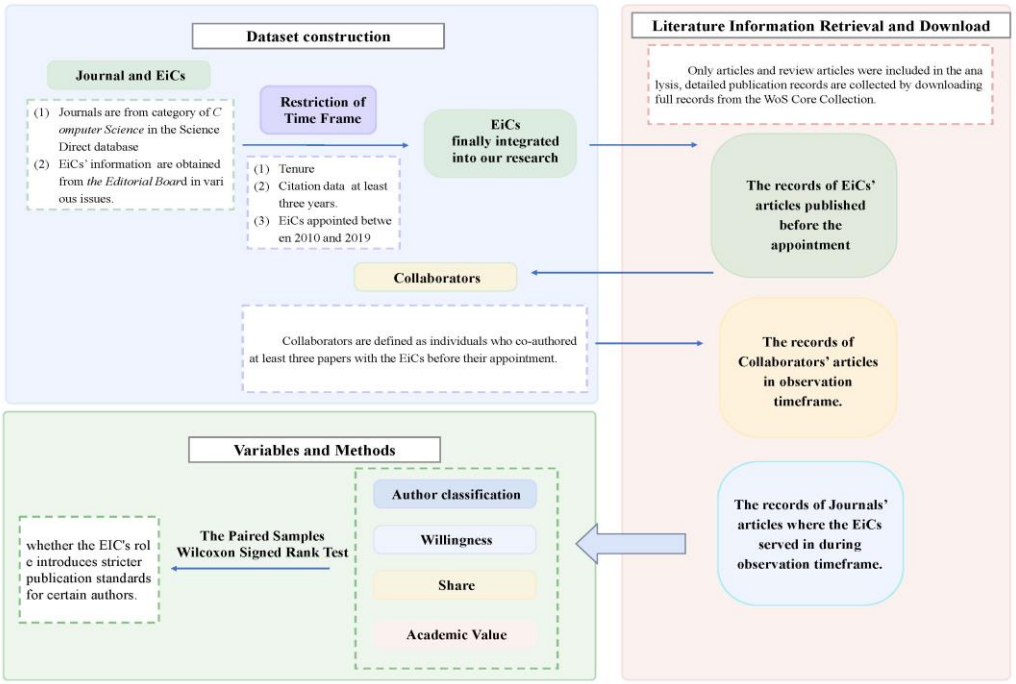


Figure 1. Research framework.

Moreover, to further enhance the comparability of the EiCs we limited the time frame to after 2010 (to reduce differences brought by time) and select EiCs from the same field (to avoid differences brought by the field, in this paper the field of computer science) for analysis. This results in a total of 48 EiCs from 40 journals.

As for collaborators, we included scholars who had at least three collaborations with the EiCs before their appointment, to ensure that the collaborators had a substantial academic relationship with the EiCs (Fu et al., 2014). The collaborations were determined based on the WoS database. The WoS interface provides author profiles and hyperlinks, which we used to count co-authorship occurrences before the EiCs' appointment. Few scholars in WoS have multiple profiles, likely due to changes in email addresses, research fields, or publication timing. We conducted manual checks using name and affiliation searches to address this issue. The process results in 603 collaborators.

After the EiCs and the collaborators are determined, their bibliographic data was also collected from the WoS database for further analysis.

Variable definition

This paper aims to analyze whether, after the appointment of the EiCs, (1) a collaborator's inclination to publish in a particular journal, (2) a collaborator's contributions to the journal, and (3) the academic value of a collaborator's papers are subject to any change.

Willingness will be used to measure a scholar's inclination to publish in a particular journal. Intuitively, for a given author, the higher the proportion of articles published in a specific journal relative to the total publications, the stronger his/her willingness

to contribute to that journal. Specifically, we defined $W_{ij} = N_{ij}/N_i$, where N_{ij} was the number of articles published by author i in journal j, and N_i was the total number of articles published by the collaborator i. *Share* will be used to measure a scholar's contributions to a specific journal. Share was defined as $S_{ij} = N_{ij}/N_j$, where N_{ij} was the number of articles published by collaborator i in journal j, and N_j was the total number of articles published in journal j.

Further, Journal Normalized Citation Impact (*JNCI*) will be used to characterize the *Academic Value* of a research paper. Academic citations, commonly used to measure influence, provide a bibliometric means of assessing academic value (note that academic value does not equate to quality, as even incomplete or imperfect papers can have academic merit). Since papers may be published in different journals and years, raw citation counts may not be directly comparable. We used the Journal Normalized Citation Impact (JNCI) to mitigate these differences. For a given paper k, $JNCI_k=c_k/E$, where c_k was the number of citations of paper k, and E was the average number of citations for papers published in the same journal and year as k.

Statistical Analysis

Based on the bibliographic data of the EiCs and the collaborators, we calculated the Willingness, Share, and Academic Value of collaborators before and after the EiCs' appointment. Changes in these indicators were analyzed using the Paired Samples Wilcoxon Signed Rank Test (if the paired sample differences do not follow a normal distribution) or the Paired Samples T-test (if normality is satisfied). The normality of the data is assessed using the Shapiro-Wilk test for sample sizes less than 50 and the Kolmogorov-Smirnov test for larger samples.

Considering authors may not contribute equally to the research presented in a paper (Hilário et al., 2023, Costas & Bordon, 2011), we classify authors as different categories for analysis. Typically, co-authors are listed in descending order of contribution, with the first author recognized for their major role. The corresponding author, who manages communication with the journal and often organizes the research, is also considered as the key contributor, even if his/her name appears last in the author list (Hu, 2009; Mattsson et al., 2011; ICMJE, 2024; Wang et al., 2013). Hence, we categorized authors into three types: Listed Author (any scholar whose name appears in the author list), Core Author (the first author or corresponding author, or both), and Other Author (authors who are listed but not as core authors). In the following analysis, we will examine the data based on these author identities.

Let W_{ij}^{before}, W_{ij}^{after}, S_{ij}^{before}, S_{ij}^{after}, V_{ij}^{before}, V_{ij}^{after} represent the Willingness, Share, and Academic Value before and after the appointment time frame. The paired samples Wilcoxon signed-rank test and the Paired Samples T-test evaluate whether there is a statistically significant difference between matched pairs of values before and after an event. Ideally, the tests capture the extent of the relative changes between these paired indicators.

From the formulas for Willingness and Share, it is evident that notable disparities may exist between a scholar's publication capacity and a journal's publication volume, potentially leading to wide variability in the distributions of these metrics. As a result, numerical changes in Willingness and Share might not accurately reflect

the true degree of change. For example, one scholar's Willingness might increase from 0.1 to 0.15, while another's shifts from 0.01 to 0.06. Although both exhibit identical absolute changes, the relative degrees of change differ significantly. A similar issue arises with Academic Value that is measured using the average *JNCI*.

To address these discrepancies, we normalized the paired values for Willingness, Share, and Academic Value, obtaining adjusted indicators for the paired test. For instance, given W_{ij}^{before} and W_{ij}^{after}, the adjusted values were calculated as follows: Adjusted $W_{ij}^{before} = W_{ij}^{before} / (W_{ij}^{before} + W_{ij}^{after})$, Adjusted $W_{ij}^{after} = W_{ij}^{after} / (W_{ij}^{before} + W_{ij}^{after})$. For the example mentioned earlier, where one scholar's Willingness increases from 0.1 to 0.15 and another's shifts from 0.01 to 0.06, the adjusted values for the first case are 0.4 and 0.6, respectively, while for the second case, they are 0.14 and 0.86. These adjusted values more accurately capture the relative degrees of change, emphasizing the disparity between the two scenarios. The same normalization is applied to S_{ij}^{before}, S_{ij}^{after}, V_{ij}^{before} and V_{ij}^{after}.

It is to be noted that, during the normalization process, there are several special cases that require additional attention, particularly when collaborators have not published any articles before and/or after the EiCs' appointment. For the indicators of Willingness and Share, if no articles are published before or after the appointment, the normalized values for W_{ij}^{before} and W_{ij}^{after} will be (0, 1) or (1, 0), which effectively capture the change of Willingness and Share. In cases where no articles are published both before and after the appointment, W_{ij}^{before} and W_{ij}^{after} will be defined as 0.5, reflecting that there has been no change in Willingness or Share. Regarding Academic Value, it is not possible to compute V_{ij} for articles that were not published. Therefore, we consider two issues: (1) For collaborators who did not publish articles in the corresponding journal before the EiCs' appointment, but did so afterward, how does the academic value of their post-appointment publications compare to the journal's average value during the same period; (2) For collaborators who have published articles both before and after the appointment, how does the academic value of their post-appointment publications compare to those published prior to the appointment.

Results

Willingness

To ensure analytical rigor, only scholars with publications during both the pre- and post-appointment periods were included in the study. This resulted in a total of 502 Listed Authors, 311 Core Authors, and 440 Other Authors being analyzed. Note that a collaborator of an EiC can be classified as either a Core Author or an Other Author, which explains why the total number of Listed Authors does not equal the sum of Core Authors and Other Authors.

The Kolmogorov-Smirnov Test was conducted on the W_{ij} values for collaborators as Listed Author, Core Author, and Other Author, and the null hypothesis of normality was rejected in all cases (p-value < 0.05). This indicated that the data do not follow a normal distribution. Therefore, the Paired Samples Wilcoxon Signed-Rank Test was applied for further analysis.

Table 1. Rank distribution of W_{ij}^{after}-W_{ij}^{before} between collaborators' W_{ij}.

Author Identity	*Rank Type*	*Case Number*	*Sum of Ranks*
Listed Author	Negative Ranks	102	9751
	Positive Ranks	86	8015
	Zero Differences	314	
Core Author	Negative Ranks	47	1925.50
	Positive Ranks	34	1395.50
	Zero Differences	230	
Other Author	Negative Ranks	74	5127.50
	Positive Ranks	68	5025.50
	Zero Differences	86	8015

Table 2. The Paired Samples Wilcoxon Signed Rank Test results of collaborators' W_{ij}.

Author Identity	*Z*	*p-value*
Listed Author	-1.195	0.232
Core Author	-1.311	0.190
Other Author	-0.109	0.914

Table 1 presents the rank distribution of differences in scholars' W_{ij} values (W_{ij}^{after}-W_{ij}^{before}) across different authorial identities, and Table 2 shows the results of the Paired Samples Wilcoxon Signed Rank Test.

Among Listed Authors, there were 102 instances of negative ranks, 86 instances of positive ranks, and 314 cases with no differences. The sum of negative ranks (9751) slightly exceeded that of positive ranks (8015). For Core Authors, 47 negative ranks and 34 positive ranks were observed, alongside 230 cases with no differences. The cumulative sum of negative ranks (1925.5) was higher than that of positive ranks (1395.5). In the case of Other Authors, 74 negative ranks, 68 positive ranks, and 298 cases with no differences were recorded. The summed negative ranks (5127.5) slightly surpassed the summed positive ranks (5025.5).

In summary, the Paired Samples Wilcoxon Signed Rank Test showed no statistically significant changes in W_{ij} values before and after the EiCs' appointment across all three categories. The p-values for Listed Authors (0.232), Core Authors (0.190), and Other Authors (0.914) exceeded the significance threshold of 0.05, indicating after the EiCs' appointment, a slight but statistically insignificant decline in scholars' inclination to publish in the journals where the EiCs served.

Share

Share refers to the proportion of collaborators' articles published in journals edited by the respective EiCs. In the calculation of S_{ij}, a total of 603 samples were included. The results of the Kolmogorov-Smirnov test (p-value<0.05) indicated that the differences in S_{ij} as Listed Author, Core Author and Other Author did not follow a normal distribution.

Table 3 presents the rank distribution of differences in S_{ij} values (S_{ij}^{after}-S_{ij}^{before}) across different authorial identities (with a different number of samples), and Table 4 summarizes the results of the Paired Samples Wilcoxon Signed Rank Test.

Table 3. Rank distribution of S_{ij}^{after}-S_{ij}^{before} between collaborators' S_{ij}.

Author Identity	*Rank Type*	*Case Number*	*Sum of Ranks*
Listed Author	Negative Ranks	125	13293
	Positive Ranks	85	8862
	Zero Differences	393	
Core Author	Negative Ranks	67	3505
	Positive Ranks	37	1955
	Zero Differences	499	
Other Author	Negative Ranks	93	7132
	Positive Ranks	71	6398
	Zero Differences	439	

Table 4. The Paired Samples Wilcoxon Signed Rank Test results of collaborators' S_{ij}.

Author Identity	*Z*	*p-value*
Listed Author	-2.589	0.010
Core Author	-2.655	0.008
Other Author	-0.633	0.527

For Listed Authors, 125 cases showed a decrease in S_{ij}, while 85 cases show an increase. The sum of negative ranks (13,293) was higher than that of positive ranks (8,862), and the test result (p-value = 0.010) indicated a statistically significant decline in Share after the EiCs' appointment.

For Core Authors, 67 cases exhibited a decrease in S_{ij}, while 37 cases displayed an increase. The sum of negative ranks (3,505) also surpassed that of positive ranks (1,955), with a p-value of 0.008 confirming a significant reduction in Share.

For Other Authors, 93 cases showed a decrease in S_{ij} and 71 cases an increase. Although the sum of negative ranks (7,132) exceeded that of positive ranks (6,398), the test result (p-value = 0.527) suggested no statistically significant change in Share for this category.

In summary, the Share of articles had significantly declined for both Listed and Core Authors following the EiCs' appointment, while no significant changes were observed for Other Authors.

Academic Value

As the situation we mentioned at subsection *Statistical Analysis*, we discussed two scenarios: (1) collaborators who published articles in the EiCs' affiliated journal after the appointment but had not published there prior to it; (2) collaborators who published articles in the same journal both before and after the appointment.

For the first scenario, we performed a descriptive analysis, with results presented in Table 5. Regardless of author identity, the mean values of *JNCI* (of collaborators

who published articles in the EiCs' affiliated journal after the appointment) were all above 1, and there were few outliers visible in Figure 1, indicating that these articles exceed the journal's average value. Although the median values were below 1, no significant differences were observed.

Table 5. Statistics of JNCIs of authors with different identities in the first scenario.

Author Identity	*mean*	*variance*	*median*	*Q*1	*Q*3
Listed Author	1.20	0.98	0.85	0.47	1.83
Core Author	1.19	0.88	0.95	0.47	1.53
Other Author	1.42	1.53	0.94	0.51	2.13

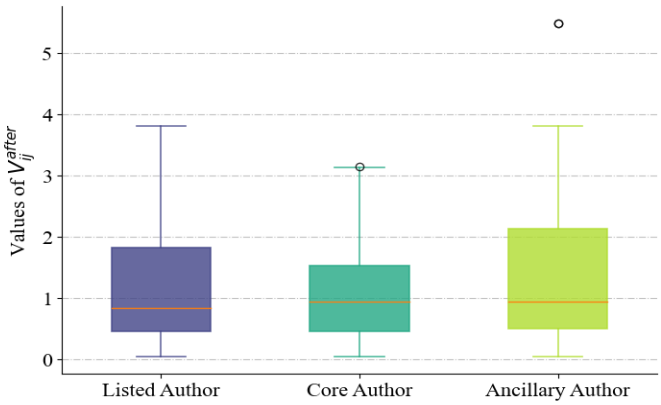


Figure 2. Boxplot of the non-optimized V_{ij}^{after} in the first scenario.

For the second scenario, we applied the Paired Samples Wilcoxon Signed Rank Test to examine changes in the value of articles published in the journals where the EiCs served. The sample included 81 Listed Authors, 26 Core Authors, and 45 Other Authors who meet the criteria. The number of Listed Authors exceeds the sum of Core Authors and Other Authors because not all Listed Authors published as Core Authors or Other Authors in both the pre- and post-appointment periods. Based on the sample size, the Kolmogorov-Smirnov test was used for the identity of Listed Author, while the Shapiro-Wilk test was applied for Core Author and Other Author. According to the test results (Listed Author, p-value = 0.94; Core Author, p-value = 0.76, Other Author, p-value = 0.30), appropriate Paired-Samples T Test was selected for further analysis.

Table 6 presents the differences of Academic Value (V_{ij}) of articles published by these scholars before and after the Editor-in-Chief's appointment, and Table 7 shows the results of Paired-Samples T-Test.

Among Listed Authors, although the sum of negative ranks exceeded that of positive ranks and the mean of (V_{ij}^{after}-V_{ij}^{before}) was less than zero, the test showed no significant changes (p-value = 0.655, greater than 0.05). For Core Authors, the sum of positive ranks exceeded the negative ranks and the mean of (V_{ij}^{after}-V_{ij}^{before}) exceeded zero, but the increase in value was not significant (p-value = 0.485, greater than 0.05).

For Other Authors, while the negative ranks slightly outnumbered the positive ranks and the mean of V_{ij}^{after}-V_{ij}^{before} with a value of 0.12, the results of the Paired Samples Wilcoxon Signed Rank Test (p-value = 0.092) indicated a decrease in value, which was statistically significant at the 90% confidence level.

Table 6. Rank distribution of V_{ij}^{after}-V_{ij}^{before} in the second scenario.

Author Identity	*Rank Type*	*Case Number*	*Involved Article Number*	*Sum of Ranks*
Listed Author	Negative Ranks	40	126	1769
	Positive Ranks	41	148	1552
	Zero Differences	0	-	
Core Author	Negative Ranks	12	41	148
	Positive Ranks	14	48	203
	Zero Differences	0	-	
Other Author	Negative Ranks	29	68	687
	Positive Ranks	16	64	348
	Zero Differences	0	-	

Table 7. The Paired-Samples t-Test results of collaborators' V_{ij} as Listed Author in the second scenario.

Author Identity	*Mean of V_{ij}^{after}-V_{ij}^{before}*	*t*	*p-value*
Listed Author	-0.02	-0.448	0.655
Core Author	0.03	0.432	0.670
Other Author	-0.12	-1.724	0.092

Discussion

The results showed that overall, after the EiCs' appointment, the collaborators' willingness to publish did not change significantly, but their publication share experienced a decline with the identities of Listed Author and Core Author. Notably, collaborators who published in the EiCs' affiliated journal for the first time after their appointment had an average article academic value exceeding the journal's average. Furthermore, for collaborators who published in the journal both before and after the EiCs' appointment, the changes in Academic Value manifested in the articles published under different identities varied: the article academic impact improved for Core Authors; while the value for Listed Authors decreased slightly, driven by a significant decline in value among Other Authors (at a 90% confidence level).

Previous research has highlighted the limited benefits collaborators of EBMs gain from their appointment. For instance, Colussi (2018) found that co-authors of EBMs don't benefit from the editor's appointment in terms of number of published papers. Similarly, Ductor and Visser (2022) noted that these collaborators seem to reap benefits that outlive the editorial term. In line with these findings, our study showed that the publication share of collaborators declined after the EiCs' appointment, suggesting that prior associations with the EiCs do not translate into preferential treatment.

However, the willingness of collaborators to submit articles remained stable. This indicates that while collaborators may not experience tangible publication benefits, they are not deterred from submitting to the correspongding EiCs' journals. This stability in Willingness can be attributed to the EiCs' aspiration to enhance the level and status of the journal, rather than harsh treatment.

When examining article value, further nuances emerge. Collaborators publishing in the EiCs' journal for the first time after their appointment exhibited article value exceeding the journal's average. This suggests that the EiCs' influence may attract submissions from high-caliber scholars, thus elevating the overall value of new contributions. By contrast, for collaborators who published both before and after the EiCs' appointment, changes in article value varied depending on their role in the authorship. Core Authors demonstrated improved article value, likely reflecting the EiCs' heightened expectations and closer scrutiny of these key contributors. Listed Authors, however, experienced a slight decline in article value, driven primarily by a significant drop among Other Authors (at a 90% confidence level). By combining the significant decline in these collaborators' Share as Core Author and the stable Share as Other Author, this trend reveals the diverse levels of responsibility and influence that different collaborator roles possess in determining the final output.

The differential treatment of collaborators can be contextualized through the lens of academic collaboration and editorial responsibility. The quality of a scholar's coauthors acts as a signal of her hidden ability and ambition the quantity and quality of one's coauthors is correlated with (Ductor et al., 2014), it can be considered that the collaborators of EiCs often possess strong academic abilities. Editors may also develop a deeper understanding of collaborators' strengths and weaknesses through prior co-authorship, making repeated collaboration a practical and cost-effective strategy for maintaining journal quality (Ductor & Visser, 2022). Consequently, EiCs may impose more stringent quality standards on submissions from trusted collaborators, especially Core Authors, to align with their responsibility to uphold journal excellence (Nourmand et al., 2024).

Finally, our findings diverge from studies emphasizing the benefits of editorial appointments. While previous research has documented advantages such as increased publication output for university colleagues (Brogaard et al., 2014) and faster handling times for papers by prior co-authors (Sarigöl et al., 2017), our results suggest a more complex dynamic. Although collaborators' submission Share decreases, the EiCs' efforts to maintain high standards ensure that the journal continues to attract quality submissions. The nuanced interplay of these factors

demonstrates how editorial appointments influence collaboration dynamics, shaping not only the distribution of publications but also their quality.

Conclusion and Limitation

This study examines the impact of EiCs' appointment on the publication behavior and academic contributions of their collaborators. Findings indicate that while collaborators' submission willingness remains stable, their publication share declines, with varying trends in article value across author roles. These findings highlight the EiCs' role in balancing editorial rigor and collaborative dynamics. However, the reliance on a single academic field and a relatively small sample size constrains broader applicability. Future research should expand the dataset to encompass journals across various disciplines, offering a more comprehensive view of EiCs-related dynamics. Exploring the effects of diverse editorial policies and collaboration patterns could provide deeper insights into how editorial leadership shapes publication practices and journal quality.

Acknowledgments

The work is supported by the Literature and Information Capacity Building Project of Chinese Academy of Science (No. E3291106).

References

Brogaard, J., Engelberg, J., & Parsons, C. A. (2014). Networks and productivity: Causal evidence from editor rotations. *Journal of Financial Economics*, 111(1), 251-270.

Colussi, T. (2018). Social ties in academia: A friend is a treasure. *The Review of Economics and Statistics*, 100(1), 45–50.

COPE. (2024). *Conflicts of interest between authors and editors*. Retrieved from https://publicationethics.org/case/conflicts-interest-issue-between-authors-and-ae

Costas, R., & Bordons, M. (2011). Do age and professional rank influence the order of authorship in scientific publications? Some evidence from a micro-level perspective. *Scientometrics*, 88(1), 145-161.

Council of Science Editors. (2024). *CSE recommendations: August 2024 edits*. Retrieved from https://cse.memberclicks.net/assets/CSE-Recommendations_Aug2024Edits_v0.pdf

Ductor, L., Fafchamps, M., Goyal, S., & Van der Leij, M. J. (2014). Social networks and research output. *Review of Economics and Statistics*, 96(5), 936-948.

Ductor, L., & Visser, B. (2022). When a coauthor joins an editorial board. *Journal of Economic Behavior & Organization*, 200, 576–595.

Fu, T. Z. J., Song, Q. Q., & Chiu, D. M. (2014). The academic social network. *Scientometrics*, 101(1), 203–239.

García, J. A., Rodriguez-Sánchez, R., & Fdez-Valdivia, J. (2015). The author–editor game. *Scientometrics*, 104, 361-380.

Hames, I. (2001). Editorial boards: realizing their potential. Learned Publishing, 14(4), 247-256.

Helgesson, G., Radun, I., Radun, J., & Nilsonne, G. (2022). Editors publishing in their own journals: A systematic review of prevalence and a discussion of normative aspects. *Learned Publishing*, 35(2), 229-240.

Hilário, C. M., Grácio, M. C. C., Martínez-Ávila, D., & Wolfram, D. (2023). Authorship order as an indicator of similarity between article discourse and author citation identity in informetrics. *Scientometrics*, 128(10), 5389-5410.

Hu, X. (2009). Loads of special authorship functions: Linear growth in the percentage of "equal first authors" and corresponding authors. *Journal of the American Society for Information Science and Technology*, 60(11), 2378-2381.

ICMJE. (2024). *Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals*. Retrieved from https://www.icmje.org/icmje-recommendations.pdf

Laband, D. N., & Piette, M. J. (1994). Favoritism versus search for good papers: Empirical evidence regarding the behavior of journal editors. *Journal of Political Economy*, 102(1), 194-203.

Liu, F., Holme, P., Chiesa, M., AlShebli, B., & Rahwan, T. (2023). Gender inequality and self-publication are common among academic editors. *Nature human behaviour*, 7(3), 353-364.

Mani, J., Makarević, J., Juengel, E., Ackermann, H., Nelson, K., Bartsch, G., ... & Blaheta, R. A. (2013). I publish in I edit?-Do editorial board members of urologic journals preferentially publish their own scientific work?. *PLoS One*, 8(12), e83709.

Mattsson, P., Sundberg, C. J., & Laget, P. (2011). Is correspondence reflected in the author position? A bibliometric study of the relation between corresponding author and byline position. *Scientometrics*, 87(1), 99-105.

Mauleón, E., Hillán, L., Moreno, L., Gómez, I., & Bordons, M. (2013). Assessing gender balance among journal authors and editorial board members. *Scientometrics*, 95, 87-114.

Nourmand, E., Swed, R., Delgado-Ruiz, R., & et al. (2024). Editors-in-chief publishing in dental journals: Concerns in self-publishing. *PLoS One*, 19(10), e0311997.

Sarigöl, E., Garcia, D., Scholtes, I., & Schweitzer, F. (2017). Quantifying the effect of editor–author relations on manuscript handling times. *Scientometrics*, 113, 609-631.

Scanff, A., Naudet, F., Cristea, I. A., Moher, D., Bishop, D. V. M., & Locher, C. (2021). A survey of biomedical journals to detect editorial bias and nepotistic behavior. *PLOS Biology*, 19(11), e3001133.

Scarlato R M, Wyburn K, Wyld M L. Is there an editorial glass ceiling? Editorial leadership in nephrology and transplantation journals: A gender-based cross-sectional analysis[J]. *Nephrology*, 2024, 29(12): 895-900.

Schubert, A. (2017). Power positions in cardiology publications. *Scientometrics*, 112(3), 1721-1743.

Tokić, B. (2017). Shared Responsibility for a Clear and Accessibly Written Scientific Paper. *Transactions of FAMENA*, 41(2), 87-104.

Trieschmann, J. S., Dennis, A. R., Northcraft, G. B., & Nieme Jr, A. W. (2000). Serving constituencies in business schools: MBA program versus research performance. *Academy of Management Journal*, 43(6), 1130-1141.

Walters, W. H. (2015). Do editorial board members in library and information science publish disproportionately in the journals for which they serve as board members?. *Journal of scholarly publishing*, 46(4), 343-354.

Wang, X., Xu, S., Wang, Z., Peng, L., & Wang, C. (2013). International scientific collaboration of China: Collaborating countries, institutions and individuals. *Scientometrics*, 95, 885-894.

Yan, Z., & Fan, K. (2024). An integrated indicator for evaluating scientific papers: considering academic impact and novelty. *Scientometrics*, 1-21.

Zdeněk, R. (2018). Editorial board self-publishing rates in Czech economic journals. *Science and engineering ethics*, 24(2), 669-682.

Zdeněk, R., & Lososová, J. (2018). An analysis of editorial board members' publication output in agricultural economics and policy journals. *Scientometrics*, 117, 563-578.

https://doi.org/10.51408/issi2025_063

How Can Citation Context Information Enrich Reference Publication Year Spectroscopy? A Case Study in Quantum Computing

Thomas Scheidsteger[1], Robin Haunschild[2], Lutz Bornmann[3]

[1]*t.scheidsteger@fkf.mpg.de*, [2]*r.haunschild@fkf.m pg.de*
IVS-CPT, Max Planck Institute for Solid State Research, Heisenbergstr. 1,
D-70569 Stuttgart (Germany)

[3]*l.bornmann@fkf.mpg.de, lutz.bornmann@gv.mpg.de*
IVS-CPT, Max Planck Institute for Solid State Research, Heisenbergstr. 1,
D-70569 Stuttgart (Germany)
Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society,
Hofgartenstr. 8, D-80539 Munich (Germany)

Abstract

Reference Publication Year Spectroscopy (RPYS) is an established bibliometric method for historical investigations of research fields based on an analysis of cited references. In this study, we propose to extend RPYS by the consideration of citation context information (CCI, location of citations in normalized sections and functions of cited papers for the citing author) to classify cited publications with respect to its specific relevance for a field. The study is based on publication metadata from the Web of Science (WoS, Clarivate). We explored the usefulness of CCI for RPYS by using exemplary publication data from the research field of quantum computing. The results show that the extension of RPYS by CCI enables more detailed analyses of cited papers (references) revealing their specific relevance for the field. The main limitation of the proposed extension is the lack of CCI data for many (older) citing publications in the WoS.

Introduction

Reference Publication Year Spectroscopy (RPYS) is an established bibliometric method for historical investigations of research fields (Bornmann & Marx, 2013; Marx, 2021). The method is not based on a times cited analysis (as most bibliometric studies), but on an analysis of cited references. In the first step of the analysis, a field-specific publication set is determined (in the Web of Science, WoS, Clarivate) such as publications dealing with quantum computing. In the second step, the cited references of these publications are analyzed with regard to the number of cited references (CRs) in each reference publication year (RPY). A plot (spectrogram) of the number of CRs against RPYs reveals peaks (RPYs including more CRs than in the neighboring RPYs) where historical roots of the field can be mostly found. RPYS can be performed by using the Java-based program Cited References Explorer (CRExplorer, https://crexplorer.net/, Thor, Bornmann, & Haunschild, 2018; Thor, Marx, Leydesdorff, & Bornmann, 2016). Since the introduction of the RPYS and the CRExplorer, more than 70 papers have been published using the method and/or the program (date of search in the WoS: January 2025). Some extensions of the initially proposed RPYS method have also been published. For example, Bornmann, Haunschild, and Marx (2023) proposed to analyze affiliation data of the CRs to

identify most referenced researchers, institutions, and countries. Ballandonne and Cersosimo (2021) introduced methods for supporting the identification of peaks in the spectrogram.

Another extension which we are going to explore in the present study is the consideration of citation context information (CCI) to classify the CRs (under the peaks) with respect to their importance further on, i.e., going beyond simple counting of occurrences. Important historical papers for a field may be extensively discussed in the field-specific citing papers which can be detected by CCI. CCI has been provided in the WoS for papers published from around 2019 onwards (Clarivate, 2022). Especially two kinds of CCI are interesting for determining importance: (1) the section of a paper according to the Introduction, Methods, Results, and Discussion (IMRaD) scheme in which a referenced paper can be found (Sollaci & Pereira, 2004), and (2) the function a referenced paper may have for the citing author (Clarivate, 2024). For example, some papers are cited only as background information in the Introduction section of a citing paper; other referenced papers are discussed in depth in the Discussion section. Since the beginning of using citations for analyzing science processes, the context of citations in publications and motivations for citing have been analyzed. Overviews of the many studies published over decades of research can be found in Bornmann and Daniel (2008) and Tahamtan and Bornmann (2019). Tahamtan and Bornmann (2019) conclude that "citing motivation is a multi-dimensional phenomenon, and scholars cite the literature for a variety of scientific and non-scientific reasons" (p. 1675).

To explore the usefulness of CCI for extending RPYS, we made a case study in the emerging research field of quantum computing. We build upon the results of Scheidsteger, Haunschild, and Ettl (2022) who performed an RPYS analysis of a publication set (citing papers) over the years 1980 to 2020 in the broader research field of quantum technology. They separated the field into four subfields: quantum metrology, quantum information, quantum communication, and quantum computing. For this case study, the subfield quantum computing (Q COMP) was selected with the largest share of papers in the whole quantum technology dataset (Scheidsteger, Haunschild, Bornmann, & Ettl, 2021). Another reason for selecting Q COMP in this study was the expectation that a large share of citing papers would be fairly recent (because of Q COMP's rapid growth). Recent publications increase the chance of available CCI in the WoS.

Using the Q COMP dataset, the following research questions have been targeted in this study:

RQ1: In which sections are cited references in the Q COMP publications primarily included and with which functions?

RQ2: Is CCI useful and suitable to enrich RPYS?

Data and Methods

Web of Science citation context data

We used an April 2024 snapshot of the WoS that includes the Science Citation Index - Expanded (SCI-E), the Social Sciences Citation Index (SSCI), the Conference

Proceedings Citation Index - Science (CPCI-S), the Conference Proceedings Citation Index - Social Science & Humanities (CPCI-SSH), and the Arts and Humanities Citation Index (AHCI). The snapshot is licensed through, and made available by, the German Kompetenznetzwerk Bibliometrie (Schmidt et al., 2024). CCI is available in the WoS since 2021 on a large scale as annotations of in-text references (from now on called "citation instances" or "citation occurrences" in this paper) that indicate why an author may have cited them. These new data are called "Enriched Cited References" by Clarivate and include (1) a numeric value between 0.0 and 1.0 for the relative position of the reference in the text of a paper, (2) the original section title and a section title that is normalized according to the IMRaD structure (Sollaci & Pereira, 2004), as well as (3) possible functions of citations. Clarivate developed a classification scheme with five functions inferred by using machine learning methods. Clarivate (2022) describes the functions as follows:

- "**Background**—previously published research that orients the current study within a scholarly area.
- **Basis**—references that report the data sets, methods, concepts, and ideas that the author is using for her work directly or on which the author bases her work.
- **Discuss**—references mentioned because the current study is going into a more detailed discussion.
- **Support**—references which the current study reports to have similar results to. This may also refer to similarities in methodology or in some cases replication of results.
- **Differ**—references which the current study reports to have differing results to. This may also refer to differences in methodology or differences in sample sizes, affecting results" (p. 1).

In the present case study, we focus on the normalized section title and the citation functions to enrich RPYS.

Table 1 shows the availability of CCI in our WoS snapshot from April 2024. We have counted for each year (1) the number of distinct papers citing a publication inside our WoS snapshot, (2) the subset of (1) with CCI, and (3) the respective percentages. We show the most recent years from 2017 onwards where the share of citing papers with CCI is above 0.1%.

Table 1. Numbers of distinct papers citing papers from our WoS snapshot in total and restricted to those with CCI as well as their respective shares over the publication years 2017 to 2024.

Publication year	*Number of citing papers*	*Number of citing papers with CCI*	*Share of citing papers with CCI*
2017	2,159,399	3,166	0.15%
2018	2,255,973	9,575	0.42%
2019	2,356,928	109,927	4.66%
2020	2,460,656	416,043	16.91%
2021	2,628,144	975,457	37.12%
2022	2,685,716	1,148,183	42.75%
2023	2,594,737	1,260,628	48.58%
2024	814,416	492,040	60.42%

Publication sets

This study is based on a dataset used by Scheidsteger, et al. (2022) who analyzed the historical roots of quantum computing using 26,650 citing papers until 2020 that had been retrieved by applying the following WoS query, see section 3.3.4 in Scheidsteger, et al. (2021): ts=("quantum hardware" OR "quantum device*" OR "quantum circuit" OR "quantum processor*" OR "quantum register*") OR ts=("quantum software" OR "quantum cod*" OR "quantum program*") OR ts=("quantum simulat*" AND (qubit* OR "quantum bit*" OR "quantum comput*") OR "quantum simulator*") OR (ts="quantum simulat*" AND wc=("quantum science technology" OR "computer science theory methods")) OR ts="quantum algorithm*" OR ts=("quantum comput*" OR "quantum supremacy" OR "quantum error correction" OR "quantum annealer" OR (quantum NEAR/2 (automata OR automaton)) OR "quantum clon* machine*").

Scheidsteger, et al. (2022) restricted the set of cited papers to those 4,459 CRs that were cited at least 25 times in the dataset of citing papers. They identified 42 seminal papers within this cited paper set. The three most-cited papers were: (1) The original idea of quantum computing from a talk on ”Simulating physics with computers” given by R. P. Feynman in 1981 was published in Feynman (1982). The paper received 1,586 citations in the dataset used by Scheidsteger, et al. (2022). (2) In a conference contribution, Peter Shor presented the first examples of quantum algorithms with a highly practical usefulness (Shor (1994) received 2,176 citations) (3) which he later was able to prove as to be polynomial-time algorithms and therefore exponentially faster than any classical algorithm (Shor, 1997). This conference contribution received 1,581 citations.

Of the 4,459 CRs in the research area of Q COMP, a subset of 3,992 could be retrieved from our WoS snapshot of April 2024. Of the 42 cited seminal papers, a subset of 38 could be retrieved from this snapshot. Since the dataset used by Scheidsteger, et al. (2022) included citing papers only until 2020, we additionally considered the more recent citing publications indexed in the WoS snapshot of April 2024. This extension is due to the recent availability of CCI in WoS as documented in Table 1. To align the selection of the additional citing papers with the focus of the original RPYS analysis by Scheidsteger, et al. (2022), we took the 3,992 cited papers of Q COMP and retrieved all citing papers with CCI from the April 2024 snapshot. The resulting dataset was restricted to Q COMP-related papers by their intersection with the outcome of the above mentioned search query in the WoS online version. Of the 3,992 CRs, 3,616 papers have been cited in 5,520 citing papers with CCI amounting to a total of 72,242 citation instances. The 38 seminal papers have been cited in 2,360 citing papers with 6,427 citation instances.

Table 2 shows the annual distribution of the 38 cited seminal papers together with the number of distinct citing papers and the total number of citation instances for which Clarivate provides CCI.

Table 2. Distribution of seminal papers of Q COMP, of their citing papers and of the CCI instances in WoS over the publication years of the cited papers.

Publication year of cited seminal papers	*Number of cited seminal papers*	*Number of citing papers with CCI*	*Number of citation instances*
1982	3	402	652
1985	1	126	188
1986	1	40	50
1989	1	30	34
1991	1	90	111
1992	2	161	263
1993	1	139	235
1994	2	539	1,010
1995	3	318	471
1996	3	173	356
1997	5	593	1,184
1998	2	136	194
1999	1	32	37
2000	1	76	135
2001	2	185	283
2003	1	133	241
2005	5	287	494
2012	2	247	399
2014	1	62	90

Table 3 shows the annual distribution of the number of papers citing the 38 seminal papers together with the number of associated citation instances. In 2017, for example, four seminal papers have been cited five times in four distinct citing papers (with CCI). Table 3 shows especially in recent years a substantial number of citation instances (starting in 2019 with nearly 100). The results thereby mirror the overall availability of CCI in the WoS as reported in Table 1.

Table 3. Distribution of papers citing the 38 seminal papers of Q COMP and of the CCI instances in WoS over the publication years of the citing papers.

Publication year of citing papers with CCI	*Number of cited seminal papers*	*Number of citing papers with CCI*	*Number of citation instances*
2008	1	1	4
2013	2	1	7
2014	1	1	2
2016	1	1	1
2017	4	4	5
2018	8	5	10
2019	27	26	98
2020	37	196	508
2021	38	498	1,379
2022	38	636	1,815
2023	38	730	1,909
2024	37	261	689

Results

Distribution of citation instances across normalized sections and citation functions

RQ1 refers to the question in which sections cited references are primarily included and with which functions. Table 4 shows the distributions of normalized sections for both cited paper sets used in this study. In both cases, the Introduction section is by far the most frequent section containing citation instances but by more than six percentage points less so for all cited papers compared to the cited seminal papers. The results also show a doubling of the shares in the Results and Methods sections, respectively, as well as an increase in the Discussion section in the case of all cited papers compared to the seminal papers.

Table 4. Distribution of normalized sections across the citation instances in papers citing the 38 seminal papers of Q COMP (total number: 6,427) and all cited papers of Q COMP (total number: 72,242).

Normalized section	*#Occ. for cited seminal papers*	*%Occ. for cited seminal papers*	*#Occ. for all cited papers*	*%Occ. for all cited papers*
Introduction	5,556	86.4	57,895	80.1
Methods	120	1.9	2,656	3.7
Results	165	2.6	3,908	5.4
Discussion	283	4.4	4,396	6.1
Not classified	303	4.7	3,387	4.7

Notes: Occ. = citation occurrence

Table 5 shows the distributions of citation functions for both sets of cited papers. In both sets, the function Background is by far the most frequent function but by about six percentage points less so for all cited papers compared to the cited seminal papers. The results are reversed for the functions Basis and Discuss: The percentages of these functions are lower for cited seminal papers than for all cited papers.

Table 5. Distribution of citation functions across the citation instances in papers citing the 38 seminal papers of Q COMP (total number 6,427) and all cited papers of Q COMP (ttal number 72,242).

Citation function	*#Occ. for cited seminal papers*	*%Occ. for cited seminal papers*	*#Occ. for all cited papers*	*%Occ. for all cited papers*
Background	5,187	80.71	53,800	74.47
Basis	454	7.06	7,081	9.80
Support	3	0.05	206	0.29
Discuss	783	12.18	11,119	15.39
Differ	0	0.0	36	0.05

Notes: Occ. = citation occurrence

Table 4 and Table 5 reveal similar trends: Compared to all cited papers, cited seminal papers tend to be more often considered as background information in introductory

sections and less often considered as sources of foundational methods (for state-of-the-art research). This observation holds at least for the years from 2019 onwards where the vast majority of CCI can be found.

RPYS enriched by normalized sections and citation functions

RQ2 concerns the usefulness of CCI for RPYS. In order to explore this question, we grouped the citation instances of the 38 cited seminal papers by their RPY, plotted the respective shares of the classes (normalized sections or citation functions) as stacked bar plots in Figure 1 and Figure 2 (resembling the RPYS spectrograms known from the RPYS analysis), and examined their peak structure (with respect to normalized sections and citation functions). The colors in the figures were chosen in a way that they visually connect normalized section titles with their most suitable counter parts in the citation functions such as Introduction with Background.

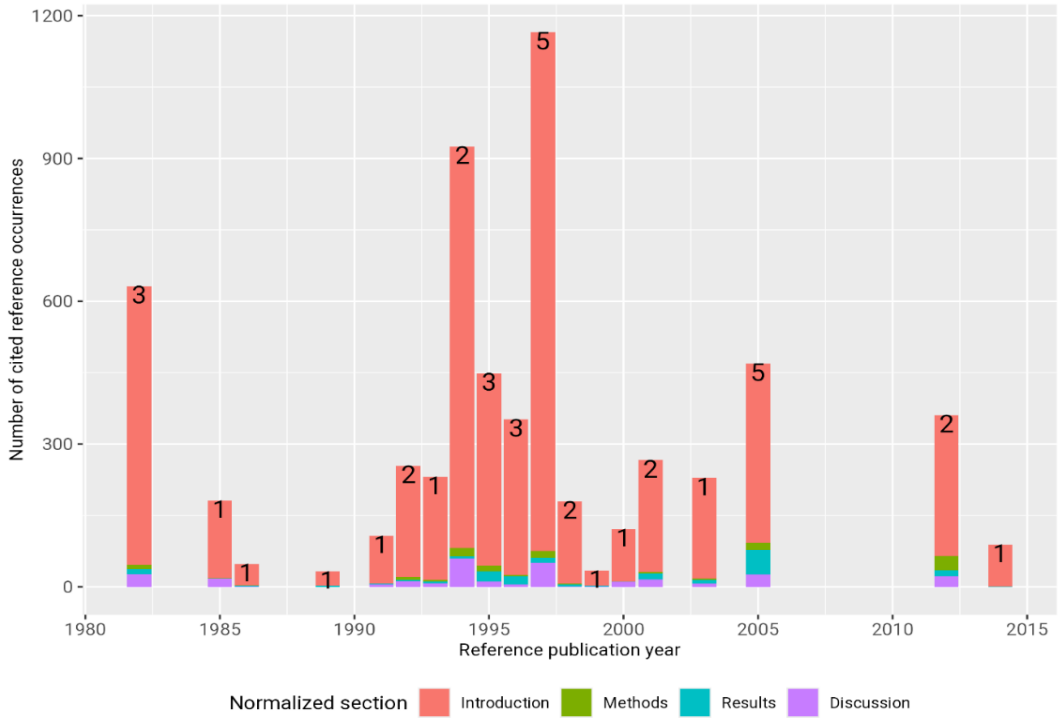


Figure 1. Annual numbers of normalized sections associated with the citation instances for the 38 cited seminal papers of Q COMP. The number mentioned in each bar is the number of cited seminal papers in the respective RPY.

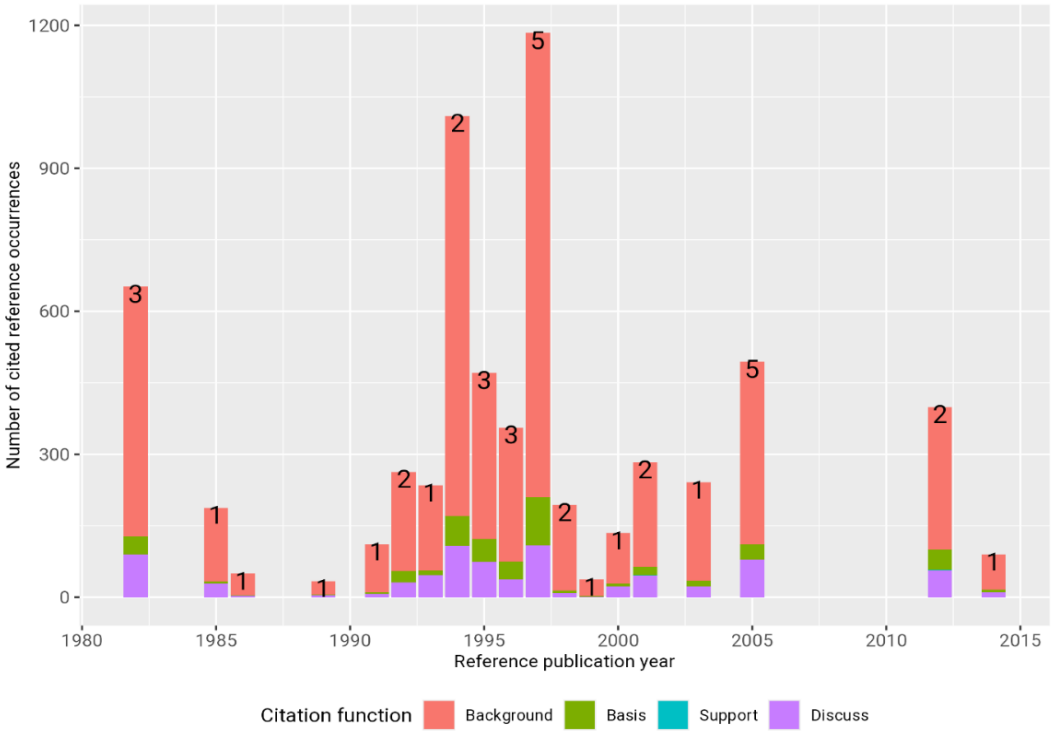


Figure 2. Annual numbers of citation functions associated with the citation instances for the 38 cited seminal papers of Q COMP. The number mentioned in each bar is the number of cited seminal papers in the respective RPY.

The RPYS method has been developed for the identification of seminal papers that are important in a certain research field. The method enables the user to identify significant peak years in the spectrogram and leads the user to those CRs that are mainly responsible for these peaks. Additional seminal papers can be detected by being long-term top-cited or even due to their outstanding absolute numbers of citations.

The results in Figure 1 and Figure 2 further specify RPYS results: The user can analyze in which sections of the citing papers and with which functions the seminal papers have been cited in certain years. If, for example, the user is interested in discussions of the cited seminal papers, the results in Figure 2 reveal that these discussions refer mainly to the cited seminal papers published around 1995. This result is confirmed by the results in Figure 1: Many seminal papers from around 1995 are cited in the Discussion section.

To exemplify the usefulness of the additional information from the CCI in this study, we tried to identify those seminal papers that have been cited due to the *methods* they offer in the realm of Q COMP. We expected that this analysis supplements the insights offered in the discussion of the cited seminal papers in Scheidsteger, et al. (2022). Figure 3 shows the numbers of citation instances associated with the section Methods and the function Basis across the RPYs of the cited seminal papers. The

Methods numbers are multiplied by three in order to facilitate comparison with the function Basis in the same plot.

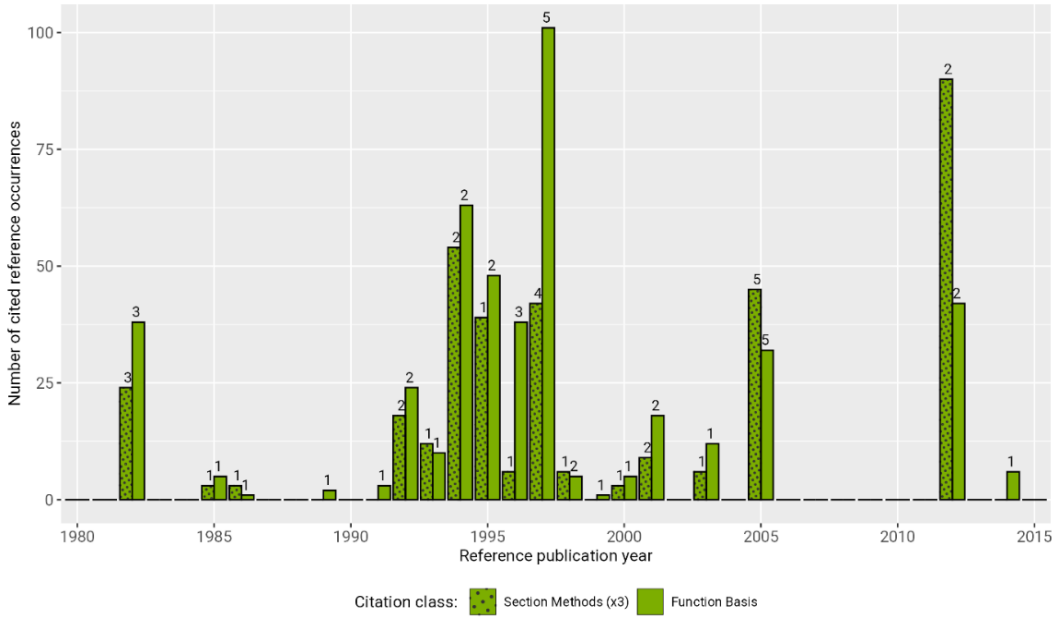


Figure 3. Annual numbers of the normalized section Methods (on the left, with dotted pattern) and the citation function Basis (on the right) associated with the citation instances for the 38 cited seminal papers of Q COMP. For better comparison, the numbers for the section Methods were multiplied by three. The number mentioned above each bar is the number of cited seminal papers in the respective RPY.

In Figure 3, we consider those RPYs with the seminal papers having the most citation instances associated with the section Methods and/or function Basis applying respective thresholds of five and 20, i.e., similar fractions of the respective maximal values. Four of these RPYs are identical with peak years of the Q COMP spectrogram in Scheidsteger, et al. (2022).

The most recent and highest Methods bar in Figure 3 is located in **2012**. Of the two cited seminal papers in this year, Fowler, Mariantoni, Martinis, and Cleland (2012) is associated with 29 of the 30 Methods citation instances. This result is confirmed by 39 of the 40 associated citation instances for the citation function Basis. Fowler, et al. (2012) provides an introduction to surface code quantum computing as one approach to construct fault-tolerant logical qubits from physical qubits. The authors of the paper intended to pave the road "towards practical large-scale quantum computation" as the title indicates. The paper is a prime example of an important methods contribution to the field.

In the next most recent year with high bars, **2005**, of 15 citation instances from five cited seminal papers, nine Methods occurrences (Basis occurrences: 16 of 32 instances) are associated with Aspuru-Guzik, Dutoi, Love, and Head-Gordon (2005). The paper presents "an efficient quantum algorithm for quantum chemical simulations of molecular energies—very much in the spirit of Feynman's original

ideas from more than 20 years before" (Scheidsteger, et al., 2022, p. 287). The authors refer to Feynman (1982) with five of the eight instances (Basis: 31 of 38 instances) in the oldest year with high bars in Figure 3, namely **1982**. Although the close connection of Feynman (1982) and Aspuru-Guzik, et al. (2005) has been identified by Scheidsteger, et al. (2022), it is confirmed by the additional analysis of section and function instances in this study.

In the year **2005**, another seminal paper contributes two of 15 instances to the Methods bar and seven of 32 instances to the Basis bar: Bravyi and Kitaev (2005) were able to significantly enlarge the error threshold in fault-tolerant quantum error-correcting schemes thus widening the operation window for quantum computing.

The year **1997** includes 14 citations to five cited seminal papers in the section Methods. The main contribution with nine citation instances stems from Shor's conference paper (Shor, 1997) assuring the quality and efficiency of the quantum algorithms he had introduced in Shor (1994). The result is confirmed by the citation function analysis: Shor (1997) accounts for 61 of 101 citation instances in this year with the highest bar for the function Basis.

The next largest contribution in **1997** (Methods: three of 14 instances; Basis: 13 of 101 instances) comes from a "Theory of quantum error-correcting codes" by Knill and Laflamme (1997). The paper plays a central role in the realization of quantum computing by stabilizing coherent states against the detrimental effect of physical noise.

A third cited seminal paper in **1997** contributes only one instance to the Methods bar, but 22 of 101 instances to the Basis bar. One reason for the few Methods occurrences is probably the fact that 17 of its Basis instances were found in papers not structured according to IMRaD, in particular not having a section Methods. Grover (1997) presents a quantum version of an unstructured search as the second practical quantum algorithm from the mid-1990s. While the algorithm "did not provide as spectacular a speed-up as Shor's algorithms, the widespread applicability of search-based methodologies has excited considerable interest in Grover's algorithm" (Nielsen & Chuang, 2010, p. 7).

Looking at cited seminal papers that are not associated with peak years in Scheidsteger, et al. (2022), we focused on the years 1992 and 1994 to 1996. The second highest Methods bar in Figure 3 refers to **1994** with a total of 18 Methods citation instances (from two cited seminal papers), 13 of which stem from Shor (1994). The author presents the first examples of quantum algorithms with a high practical value for the field—providing a first example of quantum cryptanalysis. The methodological importance of Shor (1994) is confirmed by the citation function analysis: 54 of 63 citation instances that belong to the citation function Basis go back to Shor (1994).

In **1995**, all 13 Methods instances and 47 of 48 Basis instances in Figure 3 are associated with Barenco et al. (1995). This paper provides universality proofs for certain quantum gates foundational for the construction of universal quantum circuits (Nielsen & Chuang, 2010).

The year **1996** shows a low bar with respect to the section Methods, but a high bar for the function Basis. This discrepancy can probably be explained by missing

Methods sections in citing papers (see above). Three seminal papers are contributing nearly equally to the citation instances. Steane (1996) with 13 function Basis citations used linear techniques from the classical theory of error-correcting codes to propose "Error correcting codes in quantum theory". These codes were subsequently generalized by Calderbank and Shor (1996) thereby proving that "Good quantum error-correcting codes exist". This paper gathered 15 function Basis citation instances. Despite the lack of CCI for older citing papers in WoS (see Table 3), we found that both seminal papers were cited within a section on the mathematical formalism, i.e., with explicit methodological focus, in Knill and Laflamme (1997), a seminal paper discussed above. Citations of the third seminal paper in 1996 by Bennett, DiVincenzo, Smolin, and Wootters (1996) were assigned ten times to the function Basis, and were located two times in the section Methods. In a section on the "recurrence method", this paper cites the other two seminal papers and even a preprint version of Knill and Laflamme (1997). The temporally close citation relations among these four seminal papers may point to a hot phase of methodological developments in Q COMP in the mid-1990s.

The high bar in **1992** with six citation instances from two publications is mainly (Methods: five of six instances; Basis: 23 of 24 instances) due to Deutsch and Jozsa (1992) proposing a second quantum algorithm that is proven to be faster than its classical counterpart. The latter was a classical deterministic algorithm that in the worst case would take exponentially more steps to decide the given logical problem.

Discussion

Since the introduction of RPYS, the method has been established in (professional) bibliometrics for identifying seminal papers in the history of a certain field. In this study, we investigated a possible extension of classical RPYS: the analysis of citation instances for cited publications with respect to normalized sections and citation functions in citing publications. We demonstrated the extension by using a sample dataset from the study of Scheidsteger, et al. (2022). The authors performed a classical RPYS using a publication set in quantum technology. For this case study, we focused on the subfield quantum computing (Q COMP). In order to answer RQ1, we analyzed (1) citation instances of a large set of cited papers in the subfield and, in more detail, (2) citation instances of cited seminal papers from the subfield. In both cases, citations in the Introduction section and citations classified as Background have by far the most frequent occurrences, but the citation instances of cited seminal papers show an about six percentage points higher prevalence than the citation instances of all cited papers from the subfield.

In order to answer RQ2, we exemplarily analyzed citation instances in the section Methods (2% of all CCI occurrences) and citations classified as having the function Basis (7%) for the set of cited seminal papers. For this set, we analyzed years with a lot of occurrences with respect to section Methods and/or function Basis. The analysis led to the identification of 13 cited seminal papers that can be labeled as especially important. These papers provide the methodological basis for many subsequent works in the field of Q COMP. They include, among others, the first proposals of quantum algorithms, pioneering works on quantum error correction, and

the physical implementation of fault-tolerant logical qubits. The methodological focus of these 13 seminal papers had only rudimentarily been touched in Scheidsteger, et al. (2022). Their discussion of the reasons for the importance of those seminal papers would have benefited from the inclusion of the results presented in this case study. So it seems that CCI provided by Clarivate is useful to enrich and detail results from RPYS.

What are the limitations of the analyses in this study? (1) CCI is missing for many (older) publications in the WoS. This is a main disadvantage for the application of RPYS—a method which has been especially developed for historically oriented analyses. We expect, however, that the data situation will improve constantly, since the classification of citation instances is an ongoing process at Clarivate. (2) The IMRaD section scheme is not universally applied in science; especially not in those fields that are relevant for the present study like engineering, mathematics, and computer science (Moskovitz, Harmon, & Saha, 2024). We assume thus that there are misclassifications of normalized sections in the WoS data. A closer look at the CCI in our publication sets reveals that (i) many citing papers—especially those from computer science—lack a dedicated section Methods. (ii) Several citation instances are incorrectly assigned to the section Introduction, although they can be found at a later place in the publication (manuscripts usually start with the Introduction) and point to a foundational method. (3) In this study, we focus on the normalized section title and the citation functions to enrich RPYS in a first attempt, although additional CCI is available and other CCI has been proposed in the literature. We recommend that future studies include other CCI than we did (e.g., number of papers cited to support a particular statement) and try their usefulness for the enrichment of RPYS. We would like to encourage the use of our proposed RPYS enhancement in other research fields than Q COMP. We recommend to consider especially those fields with a recent rapid growth of publications (e.g., the research field on artificial intelligence). Then, a large share of publications with CCI can be expected.

Acknowledgments

Access to WoS bibliometric data has been supported via the German Kompetenznetzwerk Bibliometrie (Competence Network for Bibliometrics, http://www.bibliometrie.info), funded by the Federal Ministry of Education and Research (grant number: 16WIK2101A). We would like to thank Clarivate for detailed descriptions of the citation context information.

References

Aspuru-Guzik, A., Dutoi, A. D., Love, P. J., & Head-Gordon, M. (2005). Simulated quantum computation of molecular energies. *Science, 309*(5741), 1704-1707. doi: 10.1126/science.1113479.

Ballandonne, M., & Cersosimo, I. (2021). A note on reference publication year spectroscopy with incomplete information. *Scientometrics*. doi: 10.1007/s11192-021-03976-1.

Barenco, A., Bennett, C. H., Cleve, R., DiVincenzo, D. P., Margolus, N., Shor, P., . . . Weinfurter, H. (1995). Elementary gates for quantum computation. *Physical Review A, 52*(5), 3457-3467. doi: 10.1103/PhysRevA.52.3457.

Bennett, C. H., DiVincenzo, D. P., Smolin, J. A., & Wootters, W. K. (1996). Mixed-state entanglement and quantum error correction. *Physical Review A, 54*(5), 3824-3851. doi: 10.1103/PhysRevA.54.3824.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45-80. doi: 10.1108/00220410810844150.

Bornmann, L., Haunschild, R., & Marx, W. (2023). Revolutions in science: The proposal of an approach for the identification of most important researchers, institutions and countries based on co-citation reference publication year spectroscopy exemplified at research on physical modelling of Earth's climate. *Journal of Information Science*, 01655515231161134. doi: 10.1177/01655515231161134.

Bornmann, L., & Marx, W. (2013). The proposal of a broadening of perspective in evaluative bibliometrics by complementing the times cited with a cited reference analysis. *Journal of Informetrics, 7*(1), 84-88. doi: 10.1016/j.joi.2012.09.003.

Bravyi, S., & Kitaev, A. (2005). Universal quantum computation with ideal Clifford gates and noisy ancillas. *Physical Review A, 71*(2), 022316. doi: 10.1103/PhysRevA.71.022316.

Calderbank, A. R., & Shor, P. W. (1996). Good quantum error-correcting codes exist. *Physical Review A, 54*(2), 1098-1105. doi: 10.1103/PhysRevA.54.1098.

Clarivate. (2022). Citation context in Web of Science. Retrieved January 30, 2025 from https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/dlm_uploads/2022/05/202205-WoS-citation-context-final.pdf

Clarivate. (2024). Enriched cited references - citation function class. Retrieved January 30, 2025 from https://webofscience.zendesk.com/hc/en-us/articles/27713957430033-Enriched-Cited-References-Citation-Function-Class

Deutsch, D., & Jozsa, R. (1992). Rapid solution of problems by quantum computation. *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences, 439*(1907), 553-558. doi: 10.1098/rspa.1992.0167.

Feynman, R. P. (1982). Simulating physics with computers. *International Journal of Theoretical Physics, 21*(6), 467-488. doi: 10.1007/BF02650179.

Fowler, A. G., Mariantoni, M., Martinis, J. M., & Cleland, A. N. (2012). Surface codes: Towards practical large-scale quantum computation. *Physical Review A, 86*(3), 032324. doi: 10.1103/PhysRevA.86.032324.

Grover, L. K. (1997). Quantum mechanics helps in searching for a needle in a haystack. *Physical Review Letters, 79*(2), 325-328. doi: 10.1103/PhysRevLett.79.325.

Knill, E., & Laflamme, R. (1997). Theory of quantum error-correcting codes. *Physical Review A, 55*(2), 900-911. doi: 10.1103/PhysRevA.55.900.

Marx, W. (2021). History of RPYS. *figshare. Conference contribution.* doi: 10.6084/m9.figshare.14910615.v1.

Moskovitz, C., Harmon, B., & Saha, S. (2024). The structure of scientific writing: An empirical analysis of recent research articles in STEM. *Journal of Technical Writing and Communication, 54*(3), 265-281. doi: 10.1177/00472816231171851.

Nielsen, M. A., & Chuang, I. L. (2010). *Quantum computation and quantum information* (10th anniversary edition ed.). Cambridge: Cambridge University Press.

Scheidsteger, T., Haunschild, R., Bornmann, L., & Ettl, C. (2021). Bibliometric analysis in the field of quantum technology. *Quantum Reports, 3*(3), 549-575. doi: 10.3390/quantum3030036.

Scheidsteger, T., Haunschild, R., & Ettl, C. (2022). Historical roots and seminal papers of Quantum Technology 2.0. *NanoEthics*. doi: 10.1007/s11569-022-00424-z.

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., . . . Stahlschmidt, S. (2024). The data infrastructure of the German Kompetenznetzwerk Bibliometrie: An enabling intermediary between raw data and analysis. doi: 10.5281/zenodo.13935407.

Shor, P. W. (1994). *Algorithms for quantum computation: Discrete logarithms and factoring.* Paper presented at the 35th Annual Symposium on Foundations of Computer Science.

Shor, P. W. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing, 26*(5), 1484. doi: 10.1137/S0097539795293172.

Sollaci, L. B., & Pereira, M. G. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association, 92*(3), 364-371.

Steane, A. M. (1996). Error Correcting Codes in Quantum Theory. *Physical Review Letters, 77*(5), 793-797. doi: 10.1103/PhysRevLett.77.793.

Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics, 121*(3), 1635–1684. doi: 10.1007/s11192-019-03243-4.

Thor, A., Bornmann, L., & Haunschild, R. (2018). CitedReferencesExplorer (CRExplorer) manual. Retrieved December 19, 2019, from https://andreas-thor.github.io/cre/manual.pdf

Thor, A., Marx, W., Leydesdorff, L., & Bornmann, L. (2016). Introducing CitedReferencesExplorer (CRExplorer): A program for reference publication year spectroscopy with cited references standardization. *Journal of Informetrics, 10*(2), 503-515. doi: 10.1016/j.joi.2016.02.005.

https://doi.org/10.51408/issi2025_086

How China and the United States Fund Artificial Intelligence? Multi-dimensional Characteristics Analysis from the Lifecycle Perspective

Hui Zhang[1], Zhe Cao[2], Ying Huang[3], Lin Zhang[4]

[1]*hui_zhang@whu.edu.cn*, [2]*caozhe@whu.edu.cn*
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)
[3]*ying.huang@whu.edu.cn*, [4]*linzhang1117@whu.edu.cn*
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)

Abstract

Funding for artificial intelligence (AI) technology research and development has ascended to a strategic priority in major global countries' scientific and technological agendas. This study explores and compares the multi-dimensional characteristics of AI-related funding projects in China and the United States (US), the global leaders in AI technology development. Specifically, it examines the characteristics of funding entities, the variety of project types, and the organization of topics across various stages of AI technology development, all contextualized within the framework of the technology lifecycle. Our results reveal that the US began funding AI technology projects earlier, and China followed a "catch-up and surpass" path. In terms of the funding agencies, while NSF, NIH, and DoD played leading roles in the US, China's main funding agencies evolved from an NSFC-centered pattern to a multi-agency balanced layout. Regarding the funding types, the US has long emphasized funding research at the applied level, which may be related to its solid technological foundation for AI development, whereas China has primarily funded research at the basic level, gradually increasing support for applied-level research as technologies mature. As for funding topics, the US funding prioritized parallel exploration of multiple topics, emphasizing interdisciplinary technological exploration and swiftly responding to technological breakthroughs to develop diverse application pathways. In contrast, China placed more emphasis on topics related to the fundamental theories and principles of machine learning and its core algorithms, reflecting a distinct evolutionary trajectory guided by national strategic priorities. These findings contribute to a deeper understanding of the differentiated developmental stages and strategic orientations of AI technologies between China and the US, serving as a reference for guiding the planning of future funding allocations.

Introduction

As the new wave of scientific and technological revolution, technological innovation has become an important tool for countries to promote economic development and enhance competitiveness. Disruptive technologies, as a critical driving force for breakthroughs at the technological frontier, not only lead industrial transformation but also reshape the global competitive landscape. Artificial intelligence (AI) stands out as a representative disruptive technology and has become a critical driving force in the current wave of technological revolution. With its remarkable capacity for innovation, disruptive impact, and far-reaching influence, AI has emerged as the

focal point of global technological competition. It plays a pivotal role in seizing key development opportunities and redefining the global industrial landscape. Numerous scientific and business organizations around the world have also listed AI technology as a representative disruptive technology, highly recognizing its value in the field of technological innovation. China and the US, as global leaders in AI technology development, play a crucial role in AI research, application, and innovation. Both countries' policies and financial investments are instrumental in driving global AI innovation and its practical deployment.

The emergence of disruptive technologies is closely related to the science and technology innovation policies and fund support of each country, and a good innovation policy can promote the emergence of disruptive technologies, thereby promoting technological innovation and industrial change and driving the development of the entire economy and society and the enhancement of competitiveness (Bin & Jieyu, 2020). As an important support for S&T innovation, fund grants drive technological progress and industrial development by supporting research projects in priority areas, carrying the national planning and strategic deployment. Given the distinct developmental trajectory of disruptive technologies, appropriate policies and projects are needed to support them at various stages of their lifecycle. This requires a strong emphasis on fundamental research and integration of fundamental research with technological innovation (Tang, Liu, Zhang, Ge, & Li, 2009; Zhao, 2022).

Current research on funding for technological innovation primarily focuses on three key aspects: the mechanisms and processes of funding, the scope of funding domains, and the evaluation of funding impacts. In studies related to the mechanisms and processes of funding, Zhao (2022) has analyzed the disruptive technology R&D and management funding systems in the US, Europe, and Japan. Based on the funding experiences of major countries globally, some studies have proposed recommendations such as establishing dedicated funding offices, forming mechanisms and funding methods for innovative technology projects, and improving the management mechanisms of science and technology projects (Sun, Zhao, & Lin, 2021; Ye, Zou, Kang, & You, 2021). Cao and Zhang (2022) used high-risk, high-reward (HRHR) research projects from typical international research institutions as an example, and explored the science and technology policy mechanisms of such research. In studies related to the layout of funding, Bai, Leng, and Liao (2017) introduced funding project data to identify frontier topics in the field of nanotechnology by using thematic clustering methods. By integrating natural language processing, text topic identification, and complex network analysis techniques, potential research frontiers were identified (Bai, Liu, & Leng, 2020). Still, Z. Q. Liu, Yue, L. X., Fang, S. (2023) have used the LDA model for funding topic detection. In studies related to the impact evaluation of funding, existing research mainly revolves around the output of scientific research results(Gao, Su, Wang, Zhai, & Pan, 2019; J. Liu & Ma, 2015; Thelwall et al., 2023). Some scholars, through citation relationships, have constructed the transformation process from fundamental research to technological innovation (Narin & Noma, 1985). For example, Du, Li, Guo, and Tang (2019) focused on the "funding-science-technology-

innovation" chain in the pharmaceutical field, they revealed the critical role of public funding in pharmaceutical innovation. Fajardo-Ortiz, Shattuck, and Hornbostel (2020) analyzed the funding landscape of major government agencies and funding organizations in the CRISPR technology field. Abadi, He, and Pecht (2020) focused on the field of artificial intelligence and compared the funding situations in China and the US. Sargent and Schwartz (2019) analyzed the development of 3D printing technology and its primary drivers.

Existing studies have provided valuable insights into the performance and distribution of funding support for technological innovation. However, they have paid limited attention to the potential differences in funding needs at various stages of technological development. These studies often fail to adequately integrate funding support with the different stages of the technological lifecycle. Therefore, this study aims to examine the evolution of the multi-dimensional characteristics of funding agency grants in China and the US from a technology lifecycle perspective, using AI technology as a case study. Specifically, this study begins by constructing the lifecycle curve of AI technology in both China and the US, clearly outlining the various stages of its development. Next, it employs machine learning methods to analyze the key characteristics of funding projects at each stage, including the evolving trends in project types and topics. Additionally, it compares these characteristics between the two countries. The main objective is to gain a deeper understanding of the funding priorities, strategies, and evolutionary trends of funding agencies in China and the US regarding disruptive technologies like AI.

Research Design

Considering the potential differences in funding needs at different stages of technological development, this study applies the technology development lifecycle framework to examine the evolving characteristics of AI funding projects in China and the United States, with a focus on how funding strategies differ at various stages of technological development. This study employs a multi-dimensional analytical framework to systematically analyze project data, focusing on three key dimensions: project entities, types, and topics. While metadata for project entities (such as funding agencies) can be directly extracted, identifying project types and analyzing topics require multi-stage data processing to obtain deeper insights. This section begins by presenting the study's analytical framework, followed by an introduction to the research data. Finally, it describes the methods used to identify project types and topics.

Analytic framework

To gain a comprehensive understanding of how funding characteristics for AI technology have evolved at different stages of development in China and the US, we first construct the lifecycle of AI technologies based on the Logistic curve. It then analyzes the characteristics of projects through three perspectives: project entity, type, and topic. The project entity perspective reveals the distribution of funding agencies, highlighting the key drivers of innovation and funding input at each stage.

The project type perspective (e.g., fundamental research, applied research, talent development, etc.) provides insights into funding strategies and resource allocation, illustrating how each country balances foundational and applied research. The project topic perspective focuses on core subjects and breakthroughs, identifying critical technological fields and research trends across various stages. These dimensions, while distinct, are interrelated. The entity characteristics address "who" is driving innovation, the type characteristics explain "how" funding supports innovation, and the topic-related characteristics reveal "which" fields and issues are prioritized. Together, they offer a comprehensive view of resource allocation and strategic priorities throughout the technology lifecycle. This analysis provides deeper insights into the funding characteristics and evolution of AI technology in China and the US. Figure 1 illustrates the research framework and methodology, which includes lifecycle curve construction and multi-dimensional analysis. The project entity analysis focuses on funding agencies, specifically examining the distribution of institutions funding AI technology in both China and the US.

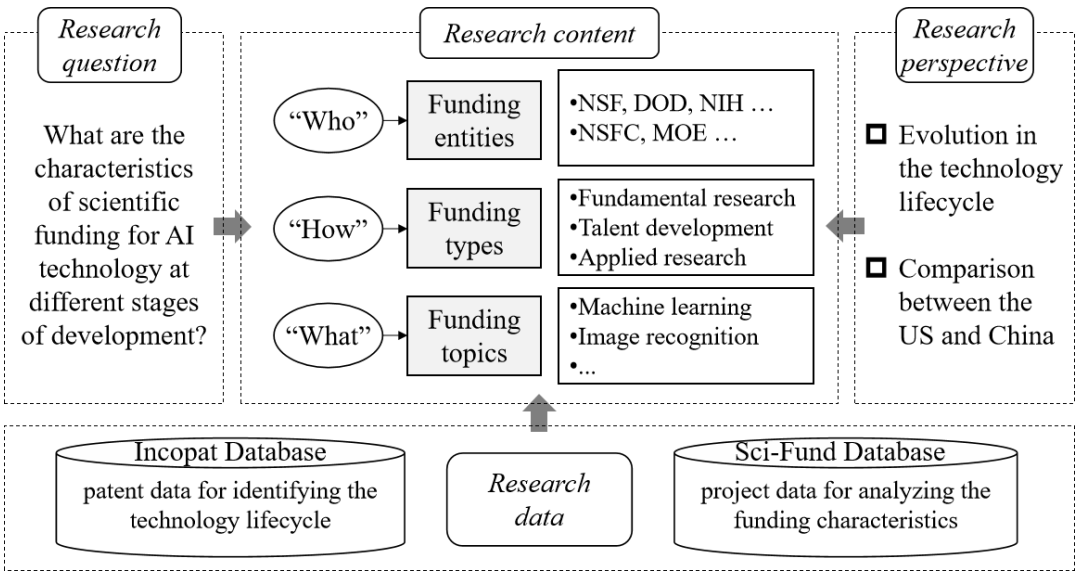


Figure 1. Analytic framework.

Data acquisition & processing

This paper selected the IncoPat and the Sci-Fund platforms as the main sources of artificial intelligence technology patent data and project data. This paper uses patent data to depict the technological lifecycle. Patent data more directly reflects the process and stages of technological innovation, making it a commonly used data source for scholars to depict technological life cycles. And IncoPat is a comprehensive global patent database covering a wide range of patent-related data, focusing on innovation trends, patent analysis, and intellectual property (IP) rights. It is one of the most widely used patent databases in China and offers data from the China National Intellectual Property Administration (CNIPA), the US Patent and Trademark Office (USPTO), along with international patent coverage from other

jurisdictions. Sci-Fund (Wanfang Sci-Fund) is a comprehensive research funding database that consolidates over 6.9 million scientific projects from nearly 20 leading nations, including China, the United States, the United Kingdom, and Japan. It integrates funding data from about 200 government agencies, national research institutions, and non-profit organizations, with continuous updates dating back to 1900. The platform relies on authoritative data sources, including direct integration with official funding repositories like the NSFC, NSF, NIH, etc. For the search of project data, keyword search is used to cover the project title and project keywords, limiting the project approval time to 2022.

The search strategy in this paper is as follows: (1) Use core keywords related to artificial intelligence technologies, such as "Artific* Intelligen*" or "AI Technolog*" as the basic search keywords; (2) Conduct separate searches for key subfields of artificial intelligence to ensure comprehensive retrieval of all project data related to AI technologies. And this paper refers to the World Intellectual Property Organization (WIPO) PATENT-SCOPE artificial intelligence index core terms and field classifications, as well as the United States Patent and Trademark Office (USPTO) field categorization of AI technology, and the finalized AI technology fields and the corresponding search keywords are shown in Table 1.

Table 1. AI technology fields and search keywords.

Classification	Subfields	Search Keywords
AI techniques	Fuzzy logic	fuzzy logic*
	Logic programming	logic* program*
	Machine learning	machine* learn*; robotic learn*; machine study*; generative AI; large language model*; general AI
	Ontology engineering	ontolog* engineer*
	Probabilistic reasoning	probabilistic reasoning
AI functional applications	Computer vision	compute* vision*; machine* vision*
	Control method	control* method*
	Distributed artificial intelligence	distribut* artific* intelligen*
	Knowledge representation and reasoning	knowledge representat* and reason*; knowledge process*; knowledge handl*
	Natural language processing	natur* language process*
	Planning and scheduling	plan* and schedul*; plan* and control*
	Predictive analytics	predict* analysis*; forecast* analysis*
	Robotics	intelligen* Robot*; smart* robot*
	Speech processing	speech process*; vioce process*
	AI hardware	AI hardware*; artific* intelligen*

		hardware*	
	Evolutionary	evolutiona*	computation*;
	computation	evolutiona*	algorithm*

There are two main steps in the data processing process.

(1) Data cleaning: Research obtains artificial intelligence patent data for technology lifecycle characterization and project data for studying project characteristics. First, the acquired patent data was deduplicated, resulting in 13,429 US patent applications and 79,302 Chinese patent applications, with annual cumulative patent application statistics. Second, text data was extracted from project titles, abstracts, and keywords, and the text data was cleaned. Missing values were manually supplemented by accessing original data sources, Chinese and US funding agency websites, and other available project information platforms to ensure data accuracy and reliability. Duplicated project data was deleted in order to reduce the interference of the noisy words in the experiments. After data processing, 28,171 Chinese-funded AI technology projects and 15,398 US-funded projects were obtained. Some project data was presented in traditional Chinese characters, requiring conversion to simplified Chinese for word segmentation processing.

(2) Text Segmentation: In this paper, we use the *jieba* library in Python for word segmentation of Chinese text in the data, and use the *spaCy* for word segmentation of English text. *Jieba* provides three different modes: precise mode, full mode and search engine mode. The precise mode is highly effective for analyzing text, as it accurately slices statements and removes redundant data, resulting in a cleaner output that avoids ambiguity and noisy words. For this reason, this paper utilizes the precise mode to process Chinese text. After conducting preliminary word segmentation, we found many meaningless stop words that do not contribute to the research topic. To address this, we created a deactivation word list to filter these words. Currently, the most commonly used Chinese stop word lists include those from the Harbin Institute of Technology, Baidu, Sichuan University Machine Intelligence Laboratory, and the Chinese stop list. This paper combines elements from these four lists to create a new stop word list, referred to as "stop_words," which contains a total of 2,462 entries. Using this stop word list, the text is further subdivided into words to achieve optimal segmentation.

Type identification and analysis of projects

This study analyzes the characteristics of funding project types for AI technology across different stages of its lifecycle. Given the lack of a direct classification system for project types, this study constructs a project classification characteristics vocabulary and employs machine learning techniques to categorize the projects into three types: fundamental research, talent development, and applied research. The project classification criteria are based on the definitions and classifications of projects in existing policies or literature. According to UNESCO, scientific research and development (R&D) activities can be divided into three categories: fundamental research, applied research, and experimental development. The Law of the People's

Republic of China on Scientific and Technological Progress also clearly states the principle that scientific and technological activities should follow, namely to "encourage fundamental research driven by applied research, and promote the integrated development of fundamental research, applied research, and achievement transformation". In research, the type of funding project is divided into four categories of personnel training, fundamental research, applied research and results of the transformation or divided into three types: fundamental research, applied research and developmental research (or developmental research)(Liang, 2023). The Department of Science and Technology (DOST) of Taiwan, China, also classifies funded projects into fundamental research, applied research, technology development, commercialization, and other types.

The identification of project types in this paper is divided into three main steps.

(1) Building project classification characteristics vocabulary: Based on the above classification of scientific activities and funded projects, this study classifies national funded projects into three types: fundamental research, talent development, and applied research. Fundamental research projects focus on the in-depth exploration of scientific theories, principles, and concepts, with the aim of advancing the development of academic disciplines and fostering knowledge innovation. Talent development projects, on the other hand, are centered around cultivating high-quality scientific and technological professionals. These initiatives aim to enhance the growth of the scientific and technological workforce through education, training, and academic exchanges. Applied research projects are typically characterized by clear practical objectives and outcomes. These projects involve activities such as technology development, system design, and engineering implementation, providing specific technical solutions to real-world challenges.

As illustrated in Table 2, a comprehensive list of project classification terms is provided. When matching project types, regular expressions are used to expand and optimize the word list, thus improving the coverage of the word list and the accuracy of matching.

Table 2. Characteristic words for the classification of Fund projects.

Project Type	Project Characteristics	Characteristic Words
Fundamental Research Projects	Focus on in-depth exploration and study of scientific theories, principles, and concepts	Theory, Mechanism, Principle, Model, Basic Science, Exploratory Research, Fundamental Research
Talent Development Projects	Focus on cultivating high-quality scientific and technological talents	Training, Education, Academic Exchange, Discipline Construction, Talent Development, Talent, Faculty, Construction

Applied Research Projects	Focusing on research on the application of scientific theories and research results in solving practical problems and promoting the transformation of scientific and technological achievements	Technology Development, Application, System Design, Engineering Implementation, Solution, Applied Research, Technology, Transfer, Industrial Cooperation, Commercialization, Business Incubation, Marketing, Achievement Transformation, Industry-Academia Cooperation

(2) Machine learning and text matching: Using the above characteristic word list for preliminary project classification, through analyzing titles, keywords, and abstracts, match the vocabulary in these texts with the characteristic word list through keyword matching to classify projects into corresponding categories. Two methods are mainly used: exact matching and fuzzy matching. For fuzzy matching, the *Levenshtein* distance algorithm is mainly used to increase matching accuracy and coverage for keywords with spelling errors or variant words (e.g., "technology transformation" and "results transformation"), through the fuzzy matching algorithm. Machine learning characteristics using labeled items of categories and training the labeled items with the help of decision tree classification models to predict the classification of unmatched items. A decision tree is a tree structure where leaf nodes represent categories or labels, and internal nodes represent characteristics. The decision tree construction process is based on the training dataset, which is divided by recursively selecting the best characteristic for optimal separation of categories. It is worth noting that some projects may matching multiple types. In such cases, this study retains the multi-type attributes to reflect its multi-dimensional characteristics.
(3) Further categorization of unsuccessfully matched projects: For projects that failed to match types, the study combines the project's program affiliation and institution for further manual classification. For example, the F32 series grants from the National Institutes of Health (NIH) primarily focus on talent development, aiming to support postdoctoral researchers' development of independent scientific research capabilities, so these projects can be classified as talent development projects. (Note: Some projects that cannot be classified into types are uniformly labeled as "unclassified" and will not be included in the subsequent analysis of project type evolution characteristics).

Topic identification and analysis of projects

The project topic analysis helps identify the technological areas and innovation directions that have received prioritized support. Projects with annotated keywords are directly assigned these keywords as their thematic representation. For projects lacking keywords, the LDA topic modeling technique is applied to identify topics from project titles and abstracts. We set K=10 and the model parameters a=0.1 and b=0.02, so as to achieve the best topic recognition effect.

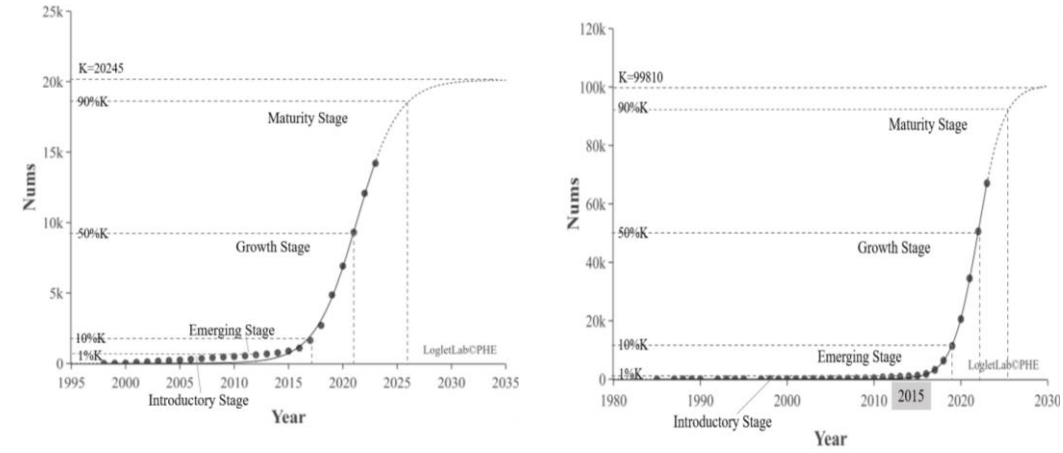
Subsequently, the study begins by analyzing the evolution of project topics, which serves as a method for detecting emerging trends. Analyzing research topics of Chinese and the US funded projects enables deep examination of topic formation, decline, strengthening, weakening, convergence, and division processes across different lifecycle stages of technological development, further characterizing strategies and features of disruptive technology development funding in both countries. Using the ItgInsight text mining tool to cluster funding topics, identify core concepts and topics of each group, and slice project topic data by year based on time series enables deeper analysis of topic changes within each time period. First, extract key topic words or phrases from each time slice; then conduct word frequency statistics on various topic words or phrases; next, select top 10 topic words by frequency each year and sort them in descending order.

Additionally, the study further calculates the strength of the co-occurrence relationship between the topic words or topic phrases in each time period, and takes the co-occurrence relationship in the previous period as the basis for measuring the strength of the relationship between the main topic words or topic phrases of the previous period and those of the next period, which is represented by a line in the graph, with a greater number of lines representing stronger co-occurrence, so as to explore the characteristics of the change of the funding topic from the perspective of evolution.

Results

AI technology lifecycle

The AI technology lifecycle serves as a central research perspective throughout the three main analyses of this study. This section, based on patent data, firstly explores the lifecycle of AI technology development in both China and the US. Due to the differences in research directions and the stages of technological development in the field of artificial intelligence between China and the United States, using patent data from both countries separately allows for precise capture of the distinct characteristics at each stage of the technological life cycle. This approach helps better understand how each country adjusts its funding strategies at different stages of AI technology development and explores the relationship between these strategies and domestic technological innovation. We use the S-curve to portray the life cycle of AI technology for auxiliary validation. The concept of the S-curve originated in 1837, first proposed by Verhulst, and is mainly classified into two types: the logistic curve and the Gompertz curve. This paper uses a Logistic model to fit the life cycle of disruptive technologies. The AI patent data from China and the US are imported into Loglet Lab4, respectively, and the patent growth data of AI technologies are fitted by the Logistic model, with the fitting results are shown in Figure 2. The goodness of fit R^2 values obtained in this paper is 0.970 and 1.000, indicating a good fitting effect.



US China

Figure 2. Lifecycle Curve of AI Technology in China and the US.

According to the model results, the US AI technology was in the introductory stage before 2012, the technological development emerging stage from 2013-2017, the growth stage from 2018-2021 and entered the maturity stage in 2022; while in China's prediction results, the AI technology is in the introductory stage before 2015, the emerging stage of technology from 2016-2019, and the period from 2020-2022 is the growth stage and entered the maturity stage in 2023.

Involvement of funding entities across AI development stages

This section begins by examining the changes in the number of funding projects in both China and the US, providing a context for understanding the distribution and dynamics of project support. It then shifts to a detailed analysis of the characteristics of the funding entities driving these projects. Statistics on the number of projects funded by China and the US each year were compiled to draw a schematic diagram, as shown in Figure 3.

As shown in Figure 3, the US started funding AI technology-related projects early, beginning in 1964. The earliest funded project was supported by NSF to the University of Kentucky Research Foundation and Case Western Reserve University Institute of Technology in 1964. Until 1985, the US remained the main sponsor of AI technology projects. Before 2006, the number of related projects funded by the US grew steadily at a relatively slow pace. From 2007 onwards, the number of AI technology projects funded by the US increased dramatically, especially after 2010, achieving an order-of-magnitude leap.

China's earliest AI technology-related projects were funded by NSFC in 1986 through a series of general programs related to AI technology. The first five institutions to receive funding were Fudan University, Peking University of Aeronautics and Astronautics, Tsinghua University, Zhejiang University, and the University of Science and Technology of China. Since then, the number of AI projects funded by China has gradually increased, especially after 2003, but at a

slightly slower rate than that of the US In 2007, China promulgated the "New Generation of Artificial Intelligence Development Plan," emphasizing the need to grasp the strategic initiative of international competition in the development of AI and the development of AI technology has entered a new period. By 2016, China's AI technology-related projects showed exponential growth, far exceeding that of the US.

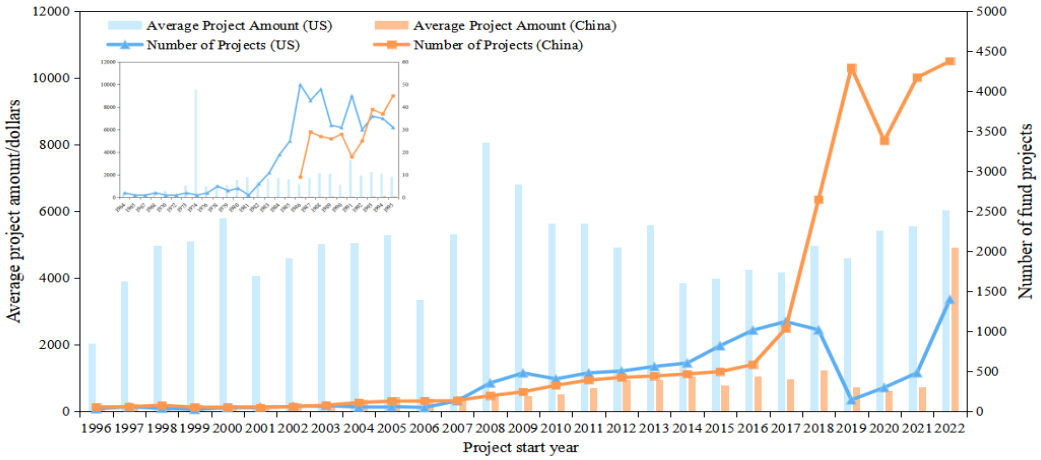


Figure 3. Trends in the number of AI-related projects funded by China and the US.

Overall, in terms of the number of funded projects from both countries, the US started to fund AI technology earlier, and was in the lead in the early stage of technology development, and the growth in the scale of funding has slowed down in recent years; while China has gradually overtaken the US in the number of grants in recent years, especially after 2016, with a faster growth rate in funding scale. In terms of funding amounts, the average amount of funding for AI technology in the US far exceeds that of China, but China has increased its funding in 2022. Combined with the lifecycle of AI technology development, China and the US in the technology development of the introductory stage of the number of grants are not high, the US took the lead in increasing the intensity of funding, in this stage of the accumulation of technology theory foundation and experience; in the technology development of the budding period, the number of grants in both countries have increased, China has entered the budding period, the rate of growth is obvious, was an explosive growth in the number of funded projects beyond the US, which reflects that China and the US for AI technology, the number of projects is more than the US. This reflects the different strategic arrangements and development grasp of AI technology between China and the US.

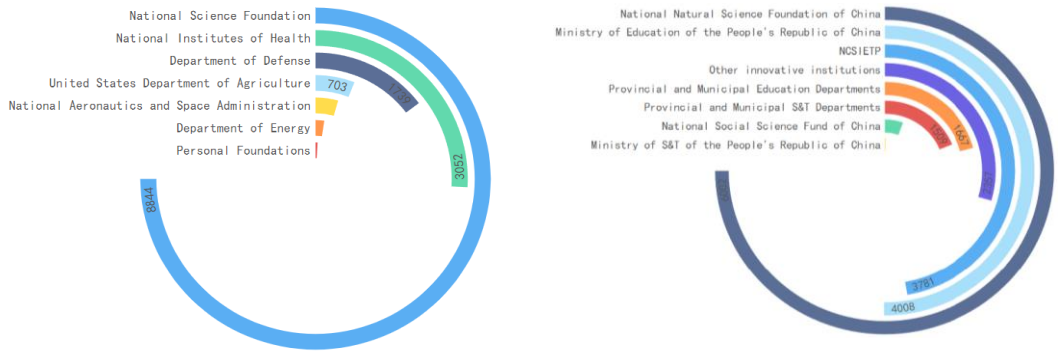
Further, the funding agencies of the two countries will be analyzed, and the strategic positioning and preferences of the US and China will be explored in promoting scientific and technological innovation, as shown in Figure 4.

It is found that the US funding for AI technology mainly comes from the National Science Foundation (NSF), the National Institutes of Health (NIH), and Department

of Defense (DoD), which fund about 88.6% of the AI projects in the US. (1) NSF primarily funds fundamental research, with universities and academic alliances being its main funding recipients, accounting for 75-80% of funding. Besides fundamental research, NSF also funds some applied research. Funded projects can be Standard Awards, which provide all funding for the entire research period within one fiscal year, or Continuing Awards, which provide project research funding incrementally over multiple years, with an average project duration of 3 years (Ma & Zhang, 2021). Notably, since 1990, NSF has implemented SGER for small-scale exploratory research projects for certain innovative research, which since 2009 has been redefined as the more targeted EAGER projects, considered "high-risk, high-reward" projects (Qiu, Jia, & Zhang, 2023). (2) The second most funded institution in the US is the National Institutes of Health (NIH), NIH-funded projects are mainly divided into Research Grants, known as R-Series Funds; Career development Awards, known as K-Series Funds and Fellowships known as F-Series Funds; and Fellowships, known as F-Series Funds. Fellowships are called F-series funds. It is worth noting that the R21 program in the R-series is a funding scheme specifically for exploratory research, with deadlines and level requirements, and is designed to encourage exploratory research by providing support for the early and conceptual stages of project development. (3) Department of Defense (DoD), with its unique Defense Advanced Research Projects Agency (DARPA), has been focusing on major breakthroughs and disruptive research projects since its inception, and has developed numerous innovations in a range of major disruptive areas such as the Internet, stealth aircraft, GPS, integrated circuits (Hao, Wang, & Li, 2015).

The top three funding agencies or major programs in China are: NSFC, the Ministry of Education of China, and the National College Students Innovation and Entrepreneurship Training Program (NCSIETP), which funded 78.7% of Chinese AI-related projects, while other funding agencies are mainly provincial and municipal Science and Technology Departments and related Science and Technology Innovation Funding Committees. NSFC is a national scientific research fund in the field of natural sciences in China, playing an important role in the development of the national innovation system, and has consistently funded fundamental research and partially applied fundamental research to support talent and team building, making remarkable contributions to the achievements and talents in China's scientific research field. In recent years, China's Ministry of Education has also gradually strengthened collaboration with the State Intellectual Property Office and other departments, and implemented a series of initiatives in conjunction with universities to strengthen cooperation and exchanges between universities, enterprises and research institutes, and to promote the transformation of scientific research results into industry. The Department of Science, Technology and Informatization, a department under the Ministry of Education, playing an important role in promoting the cultivation of scientific and technological talents and the cooperation between industry, academia and research, and has cultivated a large number of scientific and technological talents for technological development through the construction of high-level scientific and technological talent development

institutes and projects, such as key laboratories, scientific research institutes and scientific and technological innovation practice bases.

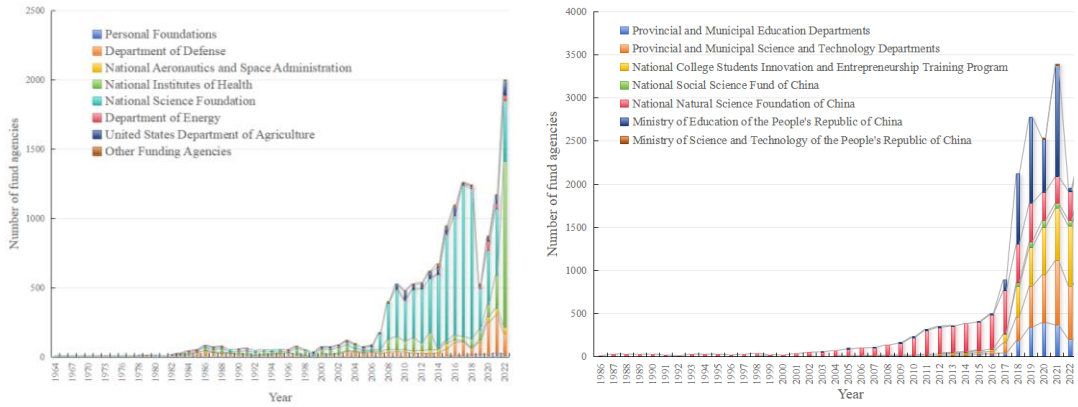


(a) US (b) China

Figure 4. Distribution of Funding Agencies in China and the US.

The evolution of funding entities in both countries, in relation to the technology development cycle, is illustrated in Figure 5. Funding from US agencies began in 1964, and in the early stage of technology development (i.e., the introductory stage), the number of funding agencies was small, primarily dominated by funding from NSF, NIH, and DoD. In 2007, there was an increase in the number of projects; by the time the technology was in its infancy, the number of funding organizations had further increased, with NSF gradually taking over as the main funder, and the US Department of Energy and the NIH gradually increasing their share of the number of funded projects. This shift may be attributed to three key factors. First, the significant advancements in artificial intelligence (AI) in the field of biomedicine likely played a crucial role. For instance, the development of AlphaFold by DeepMind in 2021(DeepMind, 2022), which solved the 50-year-old challenge of protein structure prediction, stands as one of the most groundbreaking applications of AI in science, sparking widespread attention and discussions. Second, this change is closely linked to policy initiatives from the US government. Since 2019, a series of policies have been introduced to promote the application of AI in life sciences and healthcare (COUNCIL, 2019). Notably, in 2022, the NIH released "NIH-Wide Strategic Plan" highlighting the potential of AI in health (NIH, 2022b), and the same year, NIH launched the "Bridge2AI" initiative to support the integration of multi-modal biomedical data through AI (NIH, 2022a). Finally, the COVID-19 pandemic significantly increased the demand for biomedical research and public health technologies, which may accelerate the adoption of AI in healthcare. For China's funding agencies, when in the introductory stage, the number of funding is in the stage of steady increase, the number of funding agencies is relatively small, and NSFC is the main funding agency; from 2016 when the technology development stepped into the emerging stage, with AI identified as the new engine of China's national development, the number of funding agencies and projects is explosive growth, the number of funding by the Ministry of Education and the National College

Students Innovation and Entrepreneurship Training Program of China has gradually risen and project funding peaks in 2021. This shift is closely aligned with China's policy direction. In 2018, the MOE launched the "Action Plan for AI Innovation in Higher Education" (MOE, 2018) which aimed to strengthen the development of AI disciplines and promote the integration of industry, academia, and research. This initiative may have played a significant role in increasing the number of projects funded by the MOE.



(a) US (b) China

Figure 5. The Evolution of Funding Agencies Funding AI-Related Projects in China and the US.

Distribution of funding types across AI development stages

After text analysis and type matching, the funded projects in China and the US are classified into three categories: fundamental research, talent development, and applied research. The number and distribution of each type of project in the two countries are shown in Table 3. below.

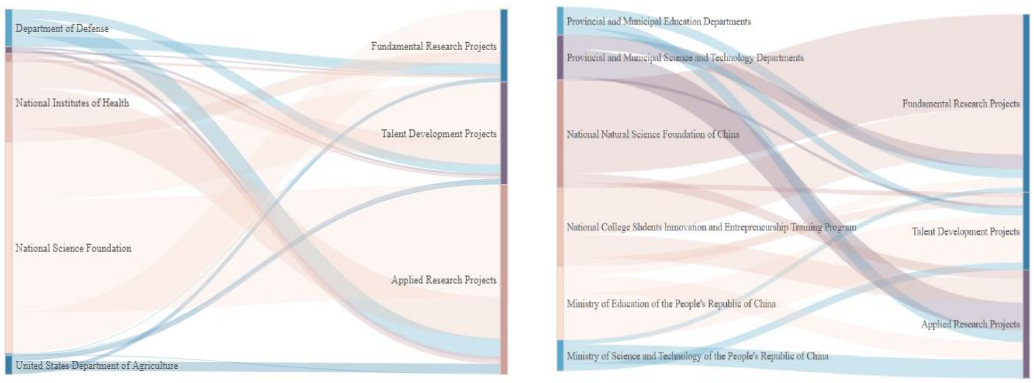
As observed from the table, China and the US show significantly different characteristics in terms of funding type. The types of projects funded by China are mainly in the category of fundamental research, which occupies more than half of the proportion, reflecting the importance China attaches to promoting the exploration of the scientific theories and principles related to AI technology. At the same time, applied research projects account for nearly 30%, indicating that China has also invested a lot of effort in promoting the practical application and commercial transformation of scientific research results. This may be related to the fact that, as mentioned earlier, the NSFC is the primary institution funding AI technology in China, with a focus on supporting fundamental research. The distribution of US funding projects for AI technology is mainly based on applied research, accounting for more than 60% of the projects, indicating that the US pays more attention to the practical application ability and market transformation potential of scientific research results.

Table 3. Distribution of funding project types in China and the US.

Classifications	The US	China
Fundamental Research Projects	28%	58%
Talent Development Projects	5%	14%
Applied Research Projects	67%	28%

In this paper, we further depict the distribution of the types of projects funded by funding agencies in China and the US, and the results are shown in Figure 6.

As observed from the figure, it can be seen that the types of agencies funding AI technology in China are more abundant, with a more balanced share of agencies, in which fundamental research projects are mainly funded by NSFC, while the Chinese Ministry of Education (MOE) and provincial and municipal education organizations are mainly funding talent development projects. Applied research projects are mainly funded by provincial and municipal science and technology organizations, but the national S&T departments are weaker in funding. In the US, NSF has invested a very high amount in both fundamental and applied research projects, occupying a major position in the funding of AI technology and highlighting its leading role in promoting the development of national AI technology in terms of scientific and technological innovation and practical application. At the same time, NIH and DoD have also shown interest in the application of AI technology in the medical and healthcare fields and national defence and the talent development projects are mainly funded by the USDA.



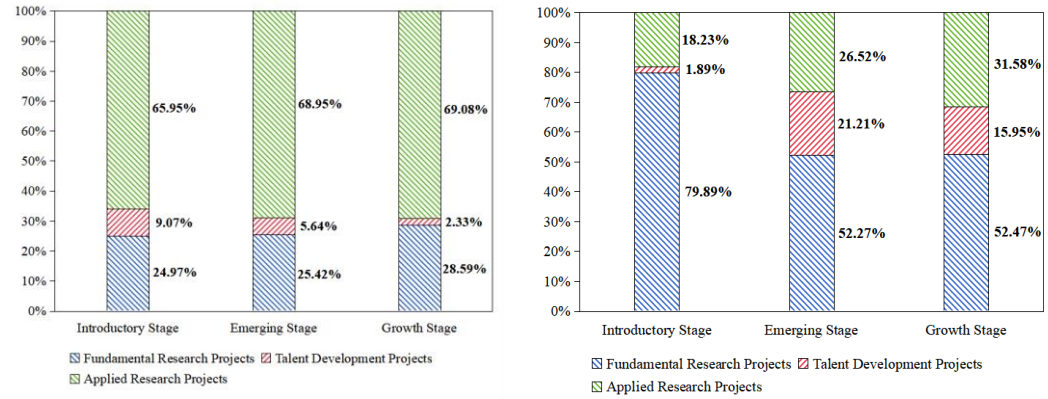
(a) US (b) China

Figure 6. Distribution of funding project types by major agencies in China and the US.

Note: The Sankey diagram shows funding agencies on the left and funding types on the right.

In order to deeply explore the funding strategies and project type characteristics of China and the US in different periods of technology development, the paper, based on the lifecycle stages of AI technology as depicted before, analyses the proportion of various types of funding projects in China and the US in different lifecycle stages. Thus, the funding bias of the two countries in different periods is reflected in Figure 7. Since the maturity stage of China and the US is incomplete, it is not counted in the statistics.

As can be seen from Figure 7, the US has led the way in applied research projects over time, with little change in the proportion of projects in each phase. Specifically, with the continuous development of AI technology, the US-funded more applied research projects and fundamental research projects aimed at expanding the boundaries of existing technologies, exploring new application areas or seeking to improve the performance and efficiency of existing technologies. The proportion of applied research projects at all stages is nearly 70%, and the proportion of talent development projects is only 2.33% as the technology enters the growth stage. China has continued to pay attention to fundamental research at all stages of the lifecycle of AI technology, and investment in fundamental research projects has always taken up a large part of the overall layout of the funding. As the technology develops into different stages, the proportion of each type of project has changed considerably. Specifically, in the technology introductory stage, nearly 80% of China's projects are funded fundamental research projects, indicating that China places particular emphasis on the exploration of basic theories in the early and middle stages of AI technology development, but only 1.89% of the projects in the category of talent development. In the emerging stage, China has increased its funding for applied research projects and talent development projects, with an increase in the ratio of 8.29% and 19.32%, respectively. However, fundamental research projects are still the main type of funding, accounting for 52.27%. When the development of technology enters the growth stage, China has further increased the funding for applied research projects, accounting for more than 30%, and has gradually put the promotion of talent development and transformation of achievements in an important position of national development. Especially since the State Council promulgated the Next-Generation Artificial Intelligence Development Plan in 2017, China has clearly put forward the ambitious goal of building a new generation of AI basic theories and key common technology systems, adhering to the application-oriented, and accelerating the commercialisation and application of AI scientific and technological achievements, China's funding for AI technology has gradually moved towards the transformation of achievements and the development of talents.



(a) US (b) China

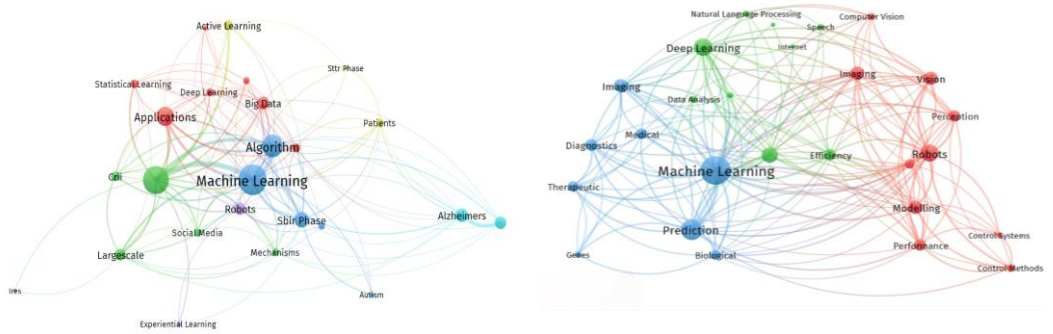
Figure 7. Distribution of funding project types at different development stages in China and the US.

Overall, the US and China have different funding types and strategies for AI technologies. The US adopts a more sustained and stable research funding strategy oriented to applied research, which is reflected in the high intensity of applied research projects at all stages of the technology's lifecycle, as well as the extensive exploration of multiple application scenarios of the technology. In contrast, China's funding is driven by fundamental research, especially under the guidance of policy, where the realization of national strategic goals is an important basis for funding. By prioritizing fundamental research projects, the Chinese government aims to strengthen the theoretical foundation of technology development and provide the necessary academic support for subsequent technological breakthroughs.

This disparity may stem from differences in the policy orientation, innovation systems, and stages of technological development between China and the US. As a latecomer in AI development, China needs to further strengthen research in foundational theories (such as algorithms and chip architectures) to reduce its reliance on Western technologies. In contrast, the US, having already established a lead in AI foundational theories (such as deep learning), is able to focus more on advancing application-driven innovations.

Evolution of funding topics across AI development stages

The word frequency of each identified project topic word was counted, and the top 30 words with the highest frequency in the two countries were extracted to construct the topic word co-occurrence network, as shown in Figure 8. This allows for the observation of the thematic focus of AI technology funding in the US and China.



(a) US (b) China

Figure 8. Thematic distribution of funded projects in China and the US.

As shown in Figure 8(a), the co-occurrence network of technology topics funded in the US reveals a tendency to support research with broad application prospects, particularly in healthcare, life sciences, and education. Among the specific funding topics, Machine Learning (ML) technologies dominate, appearing much more frequently than other technology topics. In terms of AI technology applications, topics such as "Pharmaceutical Preparations," "Biological Markers," and "Public Health" highlight the significant role of AI in biomedical research. While technologies like "Active Learning" and "Computer Simulation" demonstrate the interdisciplinary applications of AI in automation and learning processes. Figure 8(b) presents the co-occurrence network of technology topics funded in China, where "Machine Learning" also emerges as the most prominent topic, appearing far more frequently than other topics, and remains a key area of research and funding. Other frequently appearing topics include "Deep Learning" "Natural Language Processing" "Computer Vision" and "Neural Networks", which are research areas focused on the theories and principles of AI technology. In the application field of AI technology, topics such as "Internet of Things" "Robotics" and "Image Recognition", which combine AI with manufacturing and medical fields, are the most highly researched topics in AI technology application. Overall, both countries have shown a strong focus on Machine Learning, Robotics, and Natural Language Processing technologies, with ML recognized as a foundational driver of AI technology. In terms of applications, both countries emphasize the use of AI technology in healthcare, particularly with regard to China's "Healthy China 2030" initiative, which, as outlined in the 2016 policy, explicitly calls for the "development of internet-based health services" (China, 2016; Government). Additionally, while China's funding strategy places importance on integrating AI technology with national economic and social development, the US demonstrates a broader interest in interdisciplinary research.

The study further examines the evolution characteristics of funding topics. Figures 9 and 10 illustrate the evolution of funding topics in the US and China, respectively. It is worth noting that due to the longer duration of the introductory stage and the limited number of topics in the early stages of technological development, the topic

evolution graphs focus on the period from 2003 to 2022, covering the introduction, emergence, and growth stages of the technology.

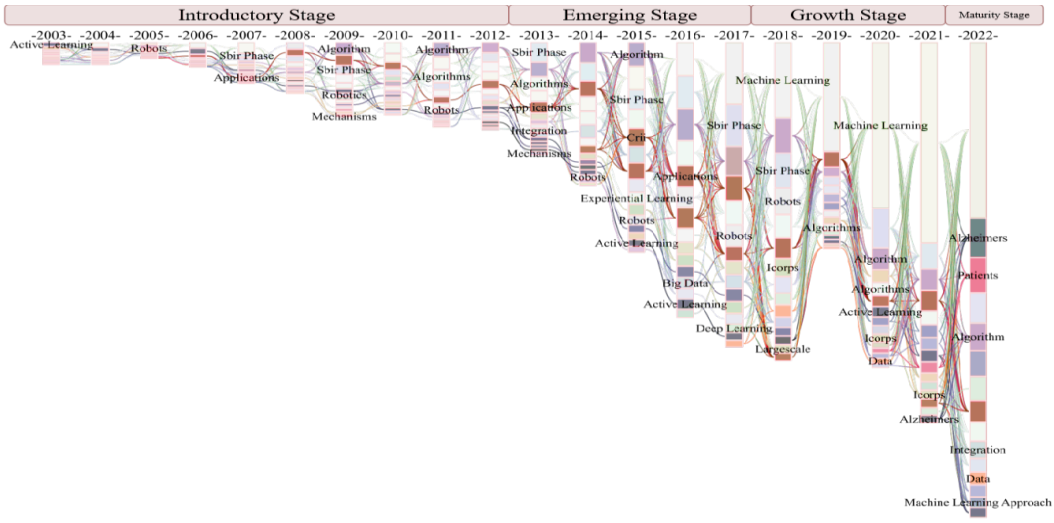


Figure 9. Evolution of topics in the US funding for AI-related projects.

It can be observed that during the early stages of AI technology development, funding topics in the US experienced fluctuating growth in both quantity and intensity. Early funding primarily focused on exploratory research and applications, covering areas such as Natural Language Processing, Robotics, Machine Learning, and related algorithms. This reflected a multi-topic, exploratory funding strategy, which was not confined to fixed research fields but encouraged cross-disciplinary innovation and diverse technological exploration. As the technology entered the emerging phase, US funding topics experienced explosive growth, quickly responding to the demands of technological development. At this stage, the scope of funding gradually extended to applications of AI across various fields such as healthcare and education. Additionally, emerging technologies, such as big data have begun to receive funding support, reflecting a broader focus on the overall technological development landscape. During this period, the US significantly increased its investment in Machine Learning, emphasizing frontier research in Machine Learning and Deep Learning. Support for other technological topics remained relatively balanced, reflecting coordinated development across multiple technology fields.

However, as the technology entered the growth phase, particularly in 2019, there was a sharp decline in the number and variety of funding topics in the AI field. This change may be attributed, on the one hand, to incomplete data collection and on the other hand, to the impact of the COVID-19 pandemic, which affected the reallocation of US government budgets and the adjustment of research priorities. In response to the pandemic, the US government shifted more resources towards urgent areas such as public health and healthcare, leading to a temporary decline in AI-related funding. Despite this, in the later stages of the pandemic, the US gradually resumed funding

for core technologies, particularly in the foundational fields of Machine Learning. At this point, funding priorities shifted towards technology applications related to public health and disease control, while new research interests emerged in technological development, further increasing support for AI application research and reflecting foresight in understanding the profound societal impact of emerging technologies.

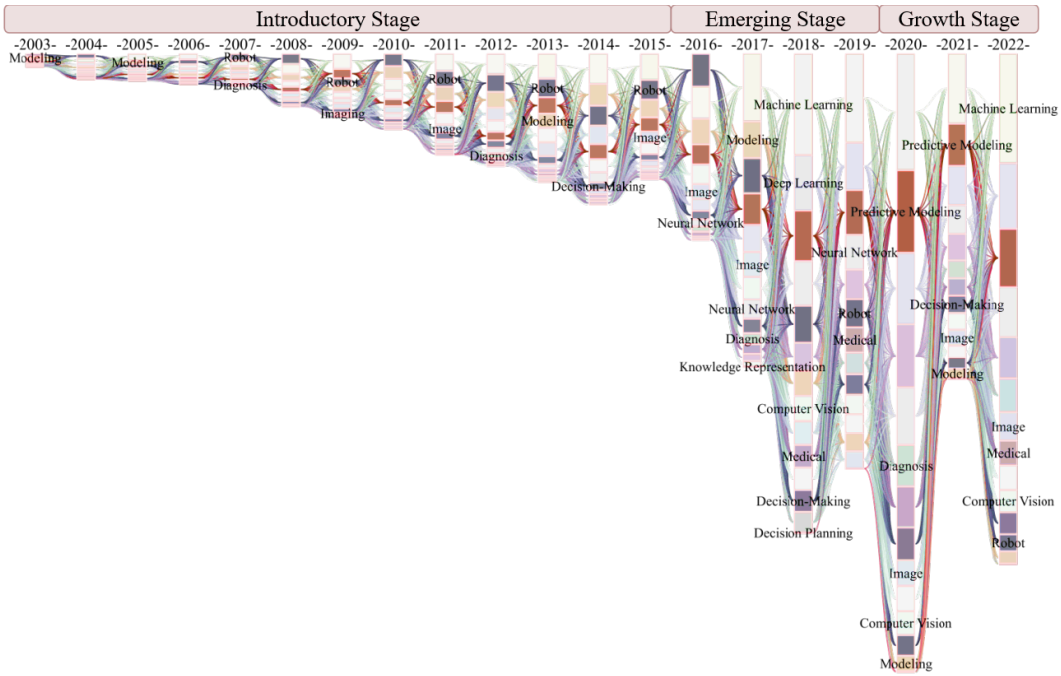


Figure 10. Evolution of topics in China funding for AI-related projects.

The evolution of funding topics in China's AI technology can be observed as follows: during the introductory stage of technology development, both the number of funding topics and the number of funded projects showed a steady upward trend. Early funding was mainly focused on fundamental research in related theories and algorithms, particularly emphasizing foundational studies in areas such as graphical learning and model construction. At the same time, exploration in fields like robotics and machine learning also received funding in the early stages. This funding during the early phase focused on laying the theoretical and algorithmic foundations for subsequent technological breakthroughs, providing solid support for driving technological innovation. As the technology entered the emerging phase, the number of funding topics and projects grew rapidly, reflecting an urgent response to the demands of technological development. During this period, China's funding still centered on the theoretical and algorithmic aspects of AI, with a particular focus on foundational research in key technologies like deep learning, natural language processing, and computer vision. At the same time, in response to the practical application needs of the technology, China began to increase investment in emerging application areas such as cloud storage, mobile robotics, and big data analytics. This

funding not only advanced fundamental research but also facilitated the transition of AI technologies into practical applications. It is noteworthy that around 2019, there was also a slight decline in the number of funding topics for AI projects.

As the technology entered the growth stage, the number of funding topics reached a new high, and funding priorities gradually shifted towards applied research. During this stage, China's funding strategy placed more emphasis on the practical applications of AI technology, particularly in areas such as the Internet of Things, big data, predictive modeling, and image recognition. During this period, funding not only promoted further technological innovation but also provided strong support for the deployment of these technologies in relevant industries. Overall, China's AI funding topics demonstrate a gradual shift from focusing on fundamental research to encompassing applied research, with funding priorities flexibly adjusted according to the practical needs of each stage of the technology lifecycle. In comparison, with the development of AI technology, both China and the US have seen a decline in the intensity of funding for AI theory and applications. US funding for AI technologies emphasizes multi-topic parallelism and technological exploration, enabling a quick response to the demands of technological breakthroughs. In the early stages of technological development, US funding placed greater emphasis on applied research and algorithm innovation, with a focus on investment in frontier technologies and emerging fields throughout the technological evolution. China's funding topics, based on machine learning, continued to increase support for key foundational technologies while exploring the application of machine learning and deep learning technologies, driving their implementation in key areas such as intelligent manufacturing and healthcare. The funding system in China is relatively centralized and stable, with a focus on fundamental research and a gradual expansion of applied research funding topics as the technology matures.

Conclusions and Discussion

This study examines AI funding strategies and characteristics in China and the US across technology lifecycle stages, revealing following conclusions: (1) In the introductory stage of AI development, US funding is predominantly led by the NSF, with an emphasis on applied research. The number of funded topics shows a fluctuating upward trend, mainly focusing on exploratory research and applied technologies. In contrast, funding in China is primarily provided by the NSFC, with approximately 80% of projects centered on fundamental research. The number of funded topics steadily increases, with research content mainly revolving theories and algorithms. (2) In the emerging stage of AI development, US funding remains primarily driven by the NSF, but the support from the DoD and NIH grows significantly. The number of funded topics experiences an explosive increase, with a focus on cutting-edge technologies such as deep learning. In China, the Ministry of Education has increased funding for AI technologies, particularly in applied research and talent development, with funding topics gradually shifting towards practical applications. (3) In the growth stage of AI development, NIH funding in the US further intensifies. In 2019, the number of funded topics sharply declined, with a shift in focus towards technology applications related to public health and

disease prevention. Meanwhile, in China, the distribution of funding agencies becomes more diversified, with funding increasingly directed towards applied research and the practical application of AI technologies. In summary, China and the US differ in the pace of AI technology development. The US started earlier and took the lead in funding earlier. NSF, NIH, and DoD played leading roles in the US With the development of technology. And US has always focused on applied research, with emphasis on multiple parallel topics. In comparison, China followed a "catch-up and surpass" path and has gradually surpassed the US, especially since 2016, with accelerated growth in funding scale. China's funding system has shifted from being primarily dominated by the NSFC to a more diversified structure with multiple agencies. And China places more emphasis on fundamental research, and is gradually expanding from fundamental research to applied research.

The AI technology funding strategies of China and the US reflect the strategic goals and policy orientations of both countries. The US emphasizes an application-oriented approach, focusing on the social impact of technologies and rapid breakthroughs. While China is driven by fundamental research, the complementary combination of fundamental research and application is gradually increasing. This disparity may stem from differences in the policy orientation, innovation systems, and stages of technological development between China and the US. As a latecomer in AI development, China needs to further strengthen research in foundational theories to reduce its reliance on Western technologies. In contrast, the US, having already established a lead in AI foundational theories, is able to focus more on advancing application-driven innovations. Although the funding priorities of the two countries are different, their respective strategies and focus adjustments reflect their deep understanding of the development of scientific and technological innovation and their forward-looking layout. In the future, China can further focus on promoting the rapid development of applied research while stabilizing basic research, enhancing cross-field cooperation, and strengthening international cooperation and global competitiveness.

Additionally, although the study analyzes the characteristics of funding projects from multiple dimensions, limitations remain, particularly in project type identification. This process relies on keyword libraries and contextual analysis, which may overlook implicit semantic relationships. Furthermore, the study is constrained by the limitations of the database used, as it does not comprehensively cover all enterprise funding data. This lack of coverage may introduce biases and affect the overall accuracy of the findings. Future research could address these limitations by incorporating more complete and diverse data sources and leveraging deep learning models, such as BERT, to better capture semantic complexity and improve data coverage.

Acknowledgments

The authors would like to acknowledge support from the National Natural Science Foundation of China (Grant Nos. 72374162, L2324105, and L2424104) and the National Laboratory Centre for Library and Information Science at Wuhan University.

References

Abadi, H. H. N., He, Z., & Pecht, M. (2020). Artificial Intelligence-Related Research Funding by the US National Science Foundation and the National Natural Science Foundation of China. *Ieee Access, 8*, 183448-183459.

Bai, R., Leng, F., & Liao, J. (2017). A Method of Detecting Research Front Based on Subjects Comparison of Multiple Data Sources. *Information Studies:Theory & Application, 40*(8), 43-48.

Bai, R., Liu, B., & Leng, F. (2020). Frontier Identification of Emerging Scientific Research Based on Multi-indicators. *Journal of the China Society for Scientific and Technical Information, 39*(7), 747-760.

Bin, W., & Jieyu, W. (2020). Analysis of Policy Requirements for Disruptive Technological Innovations:Taking Intelligent Transportation as an Example. *Journal of Technology Economics, 39*(6), 185-192.

Cao, L., & Zhang, Z. (2022). Developing Science and Technology Policies for High Risk-High Reward Research. *Bulletin of the Chinese Academy of Sciences, 37*(5), 661-673.

China, G. o. t. P. s. R. o. (2016). Healthy China 2030. Retrieved 27 January, 2025, from https://www.gov.cn/xinwen/2016-10/25/content_5124174.htm?wm=2226_2965.%202016-10-25.

COUNCIL, S. C. O. A. I. o. t. N. S. T. (2019). THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN: 2019 UPDATE. Retrieved 21 June, 2019, from https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf

DeepMind. (2022). Putting the power of AlphaFold into the world's hands. Retrieved 25 January, 2024, from https://deepmind.google/discover/blog/putting-the-power-of-alphafold-into-the-worlds-hands/

Du, J., Li, P. X., Guo, Q. Y., & Tang, X. L. (2019). Measuring the knowledge translation and convergence in pharmaceutical innovation by funding-science-technology-innovation linkages analysis. *Journal of Informetrics, 13*(1), 132-148.

Fajardo-Ortiz, D., Shattuck, A., & Hornbostel, S. (2020). Mapping the coevolution, leadership and financing of research on viral vectors, RNAi, CRISPR/Cas9 and other genomic editing technologies. *PLoS One, 15*(4).

Gao, J. P., Su, C., Wang, H. Y., Zhai, L. H., & Pan, Y. T. (2019). Research fund evaluation based on academic publication output analysis: the case of Chinese research fund evaluation. *Scientometrics, 119*(2), 959-972.

Government, C. P. s. Healthy China 2030. Retrieved January 1, 2025, from https://www.gov.cn/xinwen/2016-10/25/content_5124174.htm?wm=2226_2965.

Hao, J., Wang, H., & Li, Z. (2015). Research on DARPA's Projects Organization and Its Implications for China. *Science & Technology Progress and Policy, 32*(9), 6-9.

Liang, Q. Q. (2023). Study on NSF's Project Funding Mode for Emerging Technologies - Take the Soft Robot as an Example. *China Science & Technology Resources Review, 55*(3), 60-67.

Liu, J., & Ma, J. (2015). Progress of Basic Research in Management Science in China Based on Data from the National Natural Science Foundation. *Science and Technology Management Research, 35*(4), 249-258.

Liu, Z. Q., Yue, L. X., Fang, S. (2023). Research on Fronts Trend Prediction Method Based on the Lag of Topic Diffusion Evolution. *Information Studies:Theory & Application, 46*(6), 145-154.

Ma, W., & Zhang, S. (2021). Comparative Analysis of China and US Science Foundation Projects: Taking NSFC and NSF as Examples. *Global Science,Technology and Economy Outlook*, *36*(6), 60-72.

MOE. (2018). Action Plan for AI Innovation in Higher Education. Retrieved 26 January, 2024, from http://www.moe.gov.cn/srcsite/A16/s7062/201804/t20180410_332722.html

Narin, F., & Noma, E. (1985). IS TECHNOLOGY BECOMING SCIENCE. *Scientometrics*, *7*(3-6), 369-381.

NIH. (2022a). Bridge to Artificial Intelligence (Bridge2AI). Retrieved 25 January, 2024, from https://commonfund.nih.gov/bridge2ai

NIH. (2022b). NIH-Wide Strategic Plan for Fiscal Years 2021–2025. Retrieved 25 January, 2024, from https://www.nih.gov/sites/default/files/about-nih/strategic-plan-fy2021-2025-508.pdf

Qiu, J., Jia, Y., & Zhang, J. (2023). The Advanced Research Projects Agency for Health: a new agency in the United States and its implications for China. *Chinese Journal of Medical Management Sciences*, *13*(3), 124-128.

Sargent, J. F., & Schwartz, R. (2019). *3D printing: Overview, impacts, and the federal role*: Congressional Research Service.

Sun, Y., Zhao, B., & Lin, J. (2021). A Preliminary Study on the Whole Process Management System of Science Funds for Frontier Breakthrough Basic Research. *Science of Science and Management of S.& T.*, *42*(4), 70-82.

Tang, T. T., Liu, P., Zhang, P., Ge, F. B., & Li, M. (2009). Application of Gompertz Curve Model in the Patent Trend Forecast. *Data Analysis and Knowledge Discovery*(11), 59-63.

Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Font-Julian, C. I., et al. (2023). Is research funding always beneficial? A cross-disciplinary analysis of UK research 2014-20. *Quantitative Science Studies*, *4*(2), 501-534.

Ye, X., Zou, Q., Kang, J., & You, Y. (2021). Funding of disruptive technology:The foreign experience and China's program. *Science & Technology Review*, *39*(2), 96-103.

Zhao, Z. Y., Zhao, X. Y., Su, C., Cui, Y. W., Li, M. D. (2022). Disruptive Technology R&D and Management Funding System:Analytical Model and Foreign Practice. *Science and Technology Management Research*, *42*(17), 111-117.

https://doi.org/10.51408/issi2025_035

How Much are LLMs Changing the Language of Academic Papers?

Kayvan Kousha[1], Mike Thelwall[2]

[1]*k.kousha@wlv.ac.uk*
Statistical Cybermetrics and Research Evaluation Group, University of Wolverhampton (UK)

[2]*m.a.thelwall@sheffield.ac.uk*
Information School, University of Sheffield (UK)

Abstract

This study investigates the influence of Large Language Models (LLMs) on academic publishing with a term frequency analysis of 12 LLM-associated terms in six major scholarly databases (Scopus, WoS, PubMed, Dimensions, OpenAlex, and PMC) from 2015 to 2024. From the proportion of articles containing them, all 12 LLM-associated terms had small increases in 2023 and large increases in 2024. For example, in 2024, underscore[s/d/ing] appeared in 20% of PMC open access publications, a fivefold increase from 4% in 2022, suggesting that LLMs had influenced the language of at least 16% of PMC documents in 2024. LLM-friendly terms like delve[s/d/ing] and underscore[s/d/ing] seem to have grown partly at the expense of equivalent more traditionally academic terms like investigate[s/d/ing] and highlight[s/ed/ing]. There were disciplinary differences between the 27 Scopus broad subject categories, with underscore[s/d/ing] being more common in Environmental Science and "delve" more frequently used in Business and Humanities. There were also differences in the terms found in different parts of papers. For example, unveil[s/ed/ing] was used particularly more frequently in titles in 2024 than 2022 (0.26% vs. 0.04%), whilst underscore[s/d/ing] was more prominent in abstracts (2.5% vs. 0.21%) in Scopus. The increases may be due mainly to the use of LLMs for translation and proof reading, but imitation by researchers may result in LLM-associated terms becoming a more organic part of future academic writing, unless there is a reaction against them. Finally, since 70% of Scopus papers acknowledging ChatGPT did not use any of the 12 terms in their titles or abstracts, the influence of LLMs is probably much wider.

Introduction

Large Language Models (LLMs) like ChatGPT have the capability to help academic writing (Khalifa & Albadawy, 2024) such as editing and proofreading (Lechien et al., 2024), drafting abstracts (Gao et al., 2023; Hwang et al., 2024), creating literature reviews (Kacena et al., 2024; Margetts et al., 2024), statistical analyses (Huang et al., 2024), and even generating research hypotheses (Park et al., 2024). An Elsevier survey of researchers (n=2,284) found that about third (31%) used generative AI for research activities, and 93% found it helpful for writing and reviewing academic papers (Elsevier, 2024). A Nature survey of scientists (n=1,600) also found that almost half (47%) considered AI 'very useful' for academic tasks, with 55% believing it saves time and resources (Van Noorden & Perkel, 2023). A majority of surveyed urologists (58%, n=456) used ChatGPT for academic writing (Eppler et al., 2024) and 24% of authors in medical sciences (n=229) used LLMs for rephrasing, proofreading or translation (Salvagno et al., 2024). A survey of about 5,000

researchers found that 19% had used LLMs for the peer review process (Naddaf, 2025). Of 1,759 academic publications with ChatGPT acknowledgments, 80% mentioned language editing and proofreading or writing the manuscript and only 5% acknowledged for non-editorial research support (Kousha, 2024). However, a survey of 226 clinical researchers in 59 countries found that only 18.7% had used LLMs, mainly for grammar and formatting, and most did not acknowledge their use (Mishra et al., 2024).

Although several studies have attempted to estimate the prevalence of LLM use in academic publications, they have been limited in scope and methodology. An analysis of 2023 publications suggested that over 1% (about 60,000 papers) included LLM-associated terms (meticulously, innovatively, pivotal, intricate) (Gray, 2024). Another study found that 17.5% of Computer Science abstracts and 6.3% of Nature journal papers contained AI-modified content by using terms *realm, intricate, showcasing, pivotal* (Liang et al., 2024). In the biomedical sciences, the prevalence of LLMs terms (delves, showcasing, underscores) in PubMed abstracts rose to 10% by 2024 (Kobak et al., 2024). In dental research indexed by PubMed using terms delve, commendable, meticulous, innovative rose from 47.1 to 224.2 papers per 10,000 (Uribe & Maldupa, 2024). Using AI detection tools, a study estimated that 10% of 45,000 papers published between December 2022 and February 2023 were likely written with the help of ChatGPT (Picazo-Sanchez & Ortiz-Martin,2024). Despite these, there is a lack of subject-wide evidence from 2024, a year when a substantial fraction of authors could potentially have used ChatGPT (released November 2022) for their initial drafts, a lack of cross-database validation studies and a lack of comparisons of term frequencies in different text parts.

Research questions

This research expands on previous studies by using updated data to the end of 2024 (from 2015) and analysing the broader use of 12 LLM-associated terms across six major scholarly databases (Scopus, WoS, PubMed, Dimensions, OpenAlex, and PMC). It compares trends in the use of these terms between subjects and with other common research terms to assess changes before and after the introduction of LLMs like ChatGPT. The following research questions guide this study:

1. How has the prevalence and proportion of LLM-associated terms in academic publications changed from 2015 to 2024, and does the answer vary between major scholarly databases?
2. Are there disciplinary differences in the use of LLM-associated terms?
3. Are any LLM-associated terms particularly common in article titles or abstracts?

Methods

In this study, we investigated the potential applications of LLMs in academic writing before and after ChatGPT's November 2022 release using a range of major bibliometric databases. We searched for terms associated with LLMs in previous studies or identified through our initial tests. For the latter, we extended the list of LLM-associated terms by analysing the frequency of terms in titles and abstracts of Scopus articles in Environmental Studies. Although differences between fields are expected, Environmental Studies was selected because it is large and the identified terms were especially frequent within it.. For this we first searched for terms previously identified in related studies in titles and abstracts of Scopus articles in Environmental Studies and then identified new terms that (a) frequently co-occurred with the existing terms ($p < 0.01$, χ^2 test) and (b) had a sudden increase in frequency in 2024. We selected 12 terms to keep the analysis manageable and consistent across databases and subjects.

Table 1. lists the final terms selected for analysis in this study, along with their related sources and the queries used within the databases. Although we have no direct cause-and-effect evidence for these terms originating ever from LLMs, it seems reasonable to hypothesize that increases in their use are due to LLMs since previous research has made this assumption and the terms are general, with no obvious other source (unlike "Covid-19" or "LLM", for example).

Table 1. Identified terms potentially associated with LLM in academic publications.

Queries for terms possibly associated with LLMs	*Related source*
underscore OR underscores OR underscored OR underscoring	Kobak et al., 2024; Uribe & Maldupa, 2024
delve OR delves OR delved OR delving	Kobak et al., 2024; Uribe & Maldupa, 2024
showcasing OR showcase OR showcased OR showcases	Kobak et al., 2024; Liang et al., 2024; Uribe & Maldupa, 2024
unveil OR unveils OR unveiled OR unveiling	Uribe & Maldupa, 2024
intricate OR intricacies OR intricately	Gray, 2024; Liang et al., 2024; Uribe & Maldupa, 2024
meticulous OR meticulously	Gray, 2024; Uribe & Maldupa, 2024
pivotal	Gray, 2024; Liang et al., 2024
heighten OR heightened OR heightens OR heightening	Authors' analysis
nuanced OR nuance OR nuances	Authors' analysis
bolster OR bolstering OR bolsters OR bolstered	Authors' analysis
foster OR fostering OR fosters OR fostered	Authors' analysis
interplay OR interplays OR interplayed OR interplaying	Authors' analysis

The terms identified were searched for separately in titles, abstracts, and keywords in Scopus, WoS, and PubMed, and with unrestricted searches in three hybrid platforms that index some full text documents and some title/abstract metadata: OpenAlex, Dimensions, and PMC. The results were limited to articles, reviews, and proceedings papers published between 2015 and 2024 to analyse term usage over a decade, to guard against changes since 2022 being part of a longer-term trend unrelated to LLMs. Since the number of publications increased over time (e.g., fewer publications in 2015 than in 2024), the results were divided by the total number of publications indexed annually in each database. This approach allowed a proportional analysis of term usage, improving on previous studies that reported only raw frequency counts. All searches were conducted on 20 December 2024 to minimise the potential impact of daily increases in publications.

Results

Proportion of publications with LLM-associated terms

There were small increases in the percentage of documents containing the 12 terms in all databases in 2023 and much larger increases in nearly all cases 2024 (Figure 1). OpenAlex provides a slight anomaly, with increases in 2023 but not 2024. This might be due to OpenAlex recording the first date that it found a publication (including a preprint) rather than its formal publication date, so it may tend to be a year ahead of the other databases. In terms of other database differences, title/abstract/keyword search results for WoS and Scopus are similar but not identical, and, unsurprisingly, the highest results occur for the databases that include some full texts (PMC and Dimensions). This tends to confirm that LLMs are not only used to produce or polish article abstracts. OpenAlex is also an anomaly here, but this suggests that it indexes a low percentage of full text documents.

In 2024, *underscore[s/d/ing]* was the term most frequently used: about 20% of PMC open access publications followed by *pivotal* (15%) and a similar pattern was observed in Dimensions publications (11% and 8% respectively). Overall, the results indicate a clear and substantial overall increase in the proportion of academic publications using potentially LLM-related terms across multiple databases from 2022 onward. Figure A1 in the appendix shows the number of academic publications with LLM-associated terms across databases and years (2015-2024).

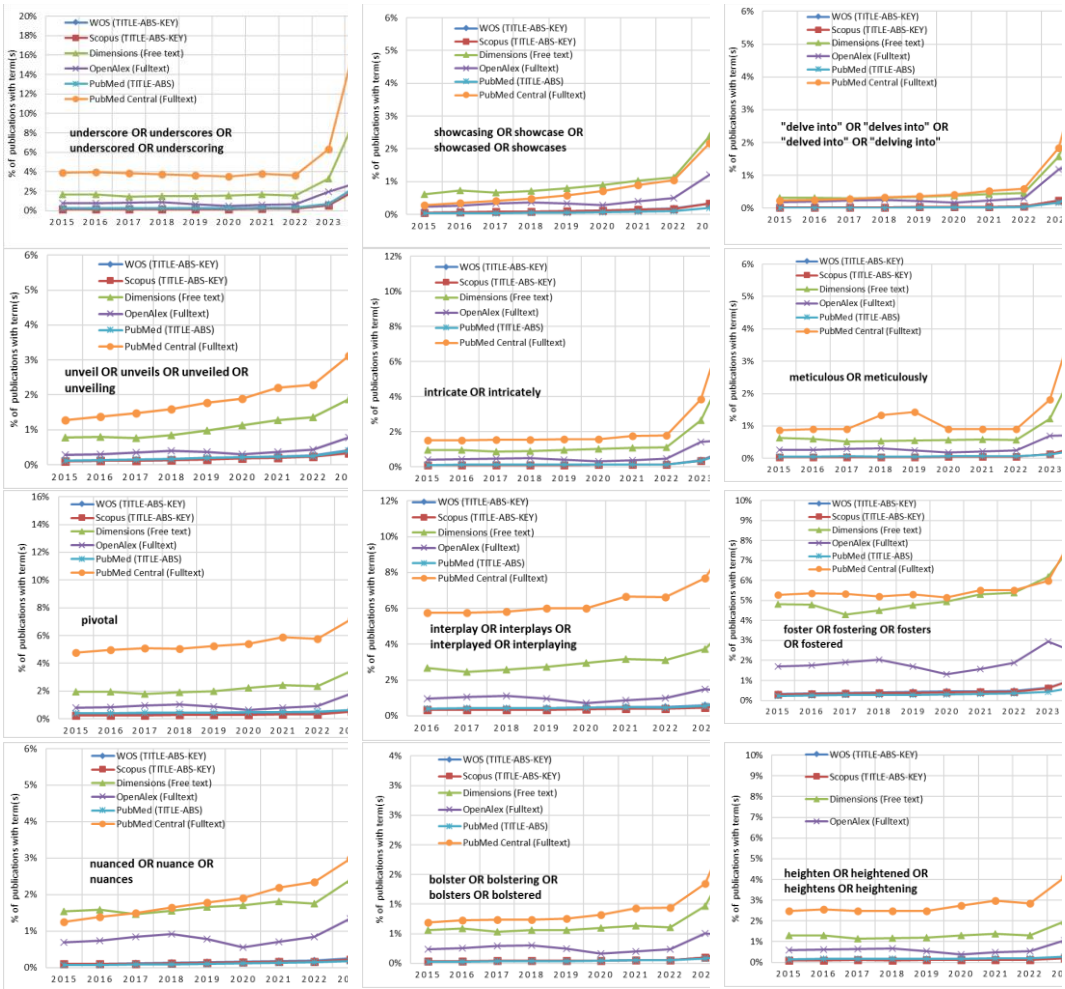


Figure 1. Percentage of 12 LLM-associated terms in academic publications in six databases.

Growth in LLM-associated terms in academic publications (2022-2024)

The terms *delve* and *underscore* had the highest growth between 2022 and 2024, with increases more than 1500% (i.e., a 15-fold increase) and 1000% in Scopus (10-fold) and WoS, respectively (Figure 2). *Intricate* and *meticulous* also experienced significant growth: above 400% in several databases. However, the terms *interplay* and *foster* had much lower increases: below 200% in several platforms. This great variability in increases may reflect a range of factors, such as their initial rarity, whether they are similar to more academic terms that they have replaced, and how often they occur in non-academic texts (where LLMs presumably learn how to use them).

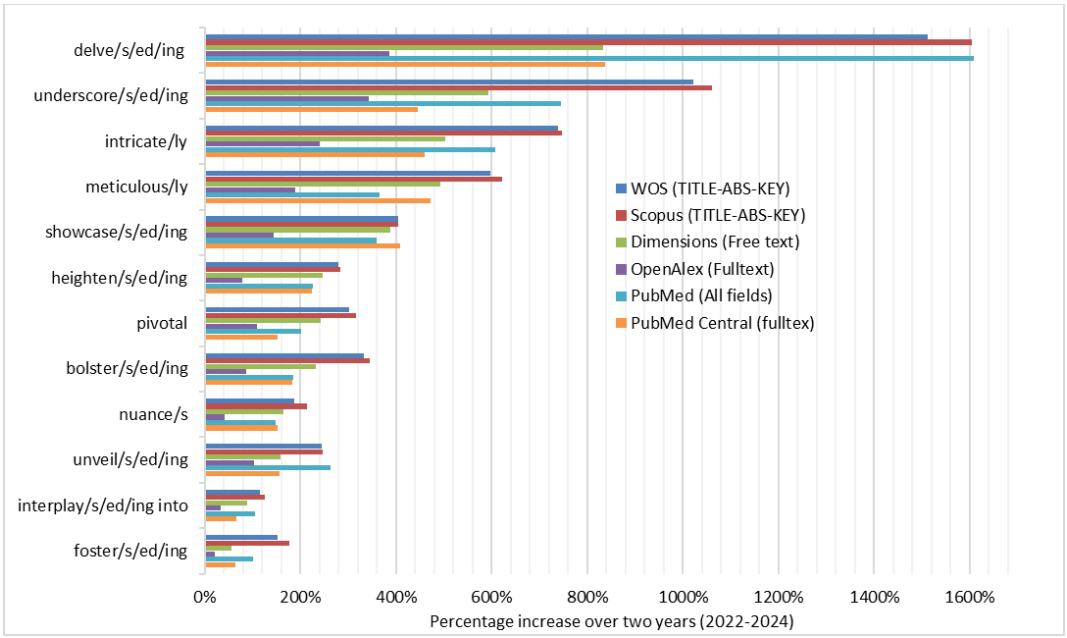


Figure 2. Percentage increase from 2022 to 2024 in the use of LLM-related terms in academic publications.

Disciplinary analysis

The percentage of LLM-associated terms in academic publications (title, abstract, or keywords) differed between Scopus subject areas in both 2022 and 2024. The term *underscore[s/d/ing]* increased dramatically in Environmental Science (0.26% to 3.84%), Business, Management, and Accounting (0.38% to 3.54%), and Economics, Econometrics, and Finance (0.35% to 3.57%) (Figure 4). Similarly, *delve[s/d/ing]* increased sharply in Business, Management, and Accounting (0.16% to 1.67%), Arts and Humanities (0.37% to 1.67%), and Economics, Econometrics, and Finance (0.12% to 1.51%) (Figures, A2 and A3, in the appendix). Hence, there seems to be some disciplinary difference in appearance of the selected terms across subjects, although this needs further investigation.

Table 2 shows the Pearson correlations between the percentage of terms in Scopus papers between 2022 and 2024 within 27 subject areas, indicating differences in their increases between disciplines. Foster[s/d/ing] (0.971), nuanced (0.966), and unveil[s/ed/ing] (0.923) have the highest correlations, suggesting a consistent increase across most subject areas and widespread usage in the titles and abstracts of academic publications. In contrast, meticulous[ly] (0.204), underscore[s/d/ing] (0.655), and bolster[s/ed/ing] (0.64) have lower correlations, indicating greater variation between subject areas, suggesting their growth may be more field-specific and could be related to research trends or discipline-specific terminology which needs further investigation.

Table 2. Pearson correlations between the percentage of LLM-associated terms in Scopus papers in 2022 against 2024 by Scopus subject. All correlations were statistically significant at the p < 0.01 level (n=27 subjects).

LLM-associated terms	*Correlation*
delve[s/d/ing]	0.807
underscore[s/d/ing]	0.655
showcase[s/d/ing]	0.744
unveil[s/ed/ing]	0.923
intricate[s/d/ing]	0.771
meticulous[ly]	0.204
heighten[s/ed/ing]	0.898
pivotal	0.677
nuance[s/d]	0.966
bolster[s/ed/ing]	0.64
foster[s/d/ing]	0.971
interplay[s/ed/ing]	0.879

The scatter plots in Figures 3 and 4 reflect a strong positive correlation between the percentage of delve[s/d/ing] and underscore[s/d/ing] in Scopus papers from 2022 to 2024 across the 27 Scopus subjects. Figure 3 shows that delve[s/d/ing] has increased consistently across most disciplines, with the highest percentages in arts & humanities, social sciences, and business. These fields have had a steady upward trend, suggesting that delve[s/d/ing] has frequently been used in abstracts or titles or recent research. In contrast, in most medical fields delve[s/d/ing] had a lower percentage increase, indicating that the term remains less commonly used in their published research.

Figure 4 also shows similar trends for underscore[s/d/ing], indicating that psychology, social sciences, environmental science, business, and economics have had the largest increases. In contrast, mathematics, physics, and dentistry have had lower percentages in using these terms. Medical subjects, such as neuroscience and medicine also showed increases, reflecting a growing use of underscore[s/d/ing] in the abstracts of Scopus papers. (see also Figure 6 below).

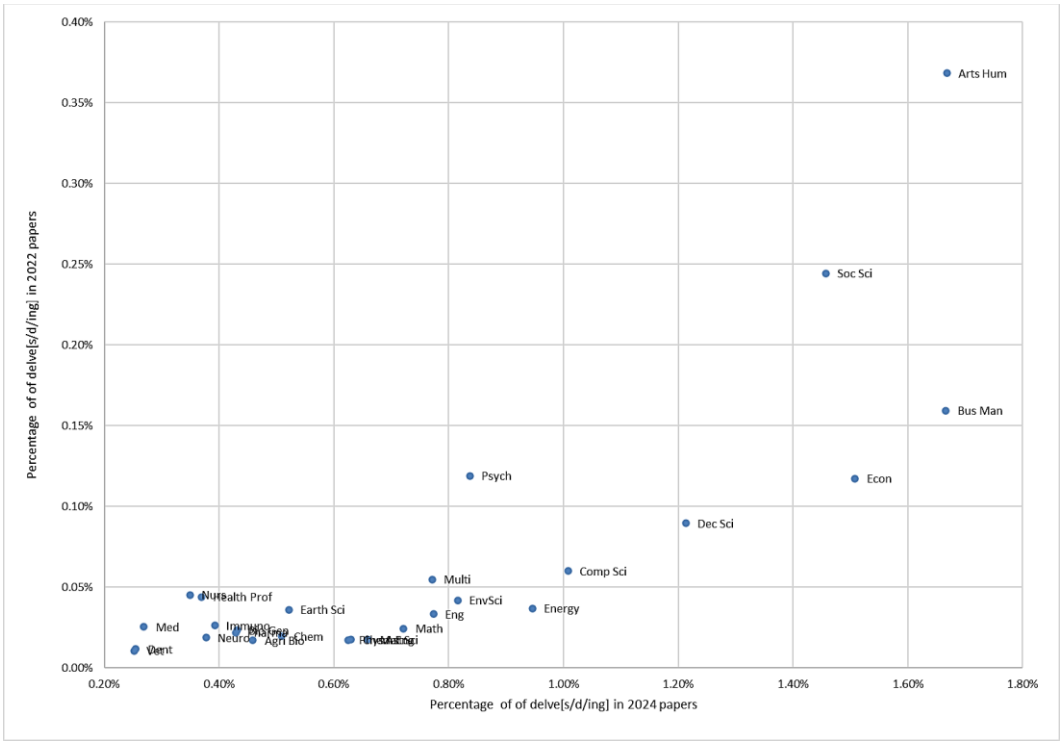


Figure 3. Scatter plot of the percentage of *delve[s/d/ing]* in Scopus papers (2022 vs 2024) across 27 subjects.

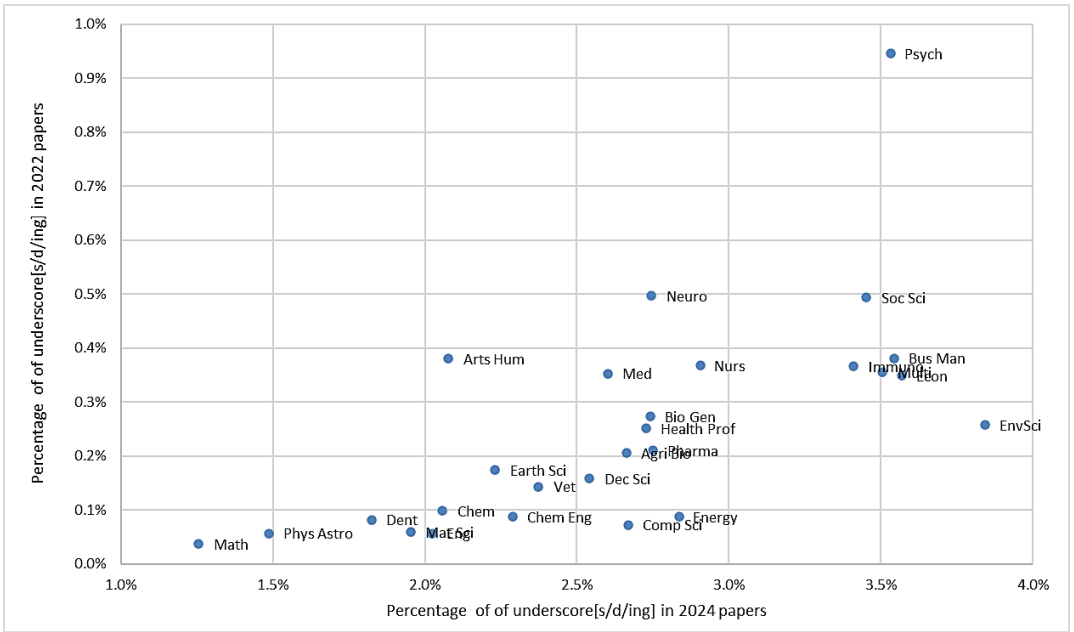


Figure 4. Scatter plot of the percentage of *underscore[s/d/ing]* in Scopus papers (2022 vs 2024) across 27 subjects.

Terms in titles and abstracts of papers

The term "unveil" was particularly common in titles in 2024 (0.26%) compared to 2022 (0.04%) and seems to be by far the most title-friendly LLM-associated term of the 12 investigated (Figure 5). In contrast, for abstracts, the term "underscore" had the biggest increase, from 0.21% in 2022 to 2.53% in 2024, and all the other terms had substantial increases (Figure 6).

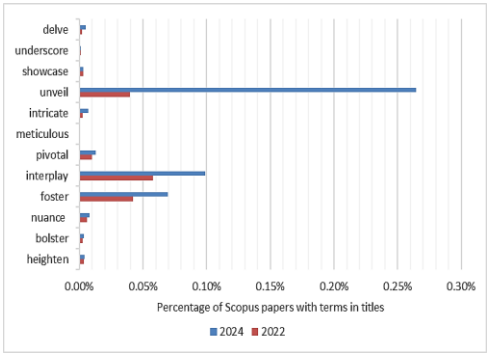


Figure 5. Percentage of Scopus publications with titles containing the selected terms (2024-2022).

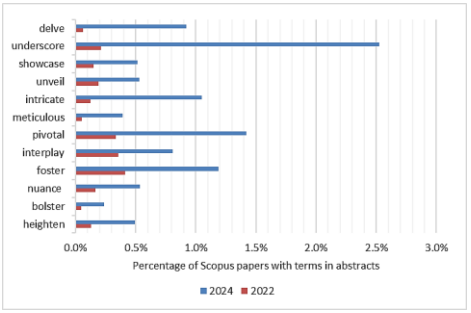


Figure 6. Percentage of Scopus publications with abstracts containing the selected terms (2024-2022).

Discussion

The results are limited by the set of 12 terms used and the six databases, and we may have overlooked terms that are used by LLMs other than ChatGPT. The results may also change in the future as LLMs evolve and if, for example, DeepSeek largely replaces existing LLMs. Our analysis is based on English-language terms and metadata (e.g., titles and abstracts) which may introduce bias. For example, non-English articles indexed with translated English abstracts could contain LLM-associated terms even if the original manuscript does not. Moreover, non-native English-speaking authors may often use LLMs for proofreading and translation to improve clarity which could influence LLM term counts.

Unlike studies that use AI detectors to identify generated text (e.g., Picazo-Sanchez & Ortiz-Martin, 2024), our approach looked at the percentage of specific vocabulary changes in publications across databases and disciplines. Although AI detectors can be used in small-scale studies, they are not practical for large-scale analyses, such as processing the abstracts of published papers across years and disciplines. Moreover, uploading academic full texts (e.g., from PMC) without authors' consent may raise ethical concerns.

Moreover, this study did not assess the average use of these terms in the full texts of publications which could provide different results compared to titles and abstracts,

where LLM-associated terms are more likely to appear only once. Hence, future studies should investigate this using large-scale data from full-text papers.

Comparing the increase of LLM with common research terms

A follow-up analysis conducted on data collected on 28 January 2025 confirmed that the use of LLM-associated terms continues to increase in frequency in the title, abstract or keywords of Scopus papers. In contrast the frequency of an ad-hoc selection of more traditional academic terms with similar meanings, used here as control terms, is relatively stable (Table 3). This supports, but does not prove, the hypothesis that LLMs are the cause of the differences rather than changes in what scientists have written about, or lengthening abstracts (which would make all terms more common).

Table 3. Percentage increase in common vs. LLM-associated terms in Scopus papers (2022–2024).

Common term	2022 (%)	2024 (%)	**% Increase**	**LLM term**	2022 (%)	2024 (%)	**% Increase (2022-2024)**
investigate	17.83	19.41	**8.90%**	delve	0.07	1.05	**1360%**
highlight	5.08	9.43	**85.68%**	underscore	0.25	2.87	**1062%**
demonstrate	14.07	20.35	**44.72%**	showcase	0.20	0.99	**395%**
reveal	12.03	16.02	**33.25%**	unveil	0.26	0.88	**235%**
complex	10.02	11.78	**17.63%**	intricate	0.14	1.20	**727%**
precise	1.74	2.93	**67.94%**	meticulous	0.06	0.45	**611%**
critical	6.45	7.99	**23.92%**	pivotal	0.40	1.62	**308%**
enhance	9.56	18.76	**96.25%**	heighten	0.15	0.57	**273%**
detail	4.16	4.37	**4.94%**	nuanced	0.20	0.61	**210%**
strengthen	1.48	1.65	**11.42%**	bolster	0.06	0.27	**361%**
promote	5.87	6.99	**19.17%**	foster	0.50	1.40	**177%**
interaction	8.25	9.06	**9.89%**	interplay	0.45	0.99	**119%**

Are academics reviewing LLM-generated texts?

The extent to which LLMs like ChatGPT are used in academic writing (e.g., minor grammatical edits, spell checking, or fully drafting sections or abstracts) requires further qualitative and quantitative investigation. However, out of 1,540 academic papers with ChatGPT acknowledgments related to manuscript editing and production (see data from Kousha, 2024), about a third (31%) included one or more of the 12 LLM-associated terms in their titles or abstracts (e.g., underscore[s/d/ing] (7.3%), pivotal (4.2%), and intricate[s/d/ing] (3.7%). Since 69% did not include any of these terms, the highest of the results above (a 16% increase for underscore)

probably underestimate the prevalence of LLM support for academic writing: the real figure may be at least triple (1/0.31=3.23) the maximum reported here. If LLMs are widely used for editing, common phrases like "underscore" or "delve" may become even more popular in academic writing in future.

Conclusions

In answer to the first research question, the findings show a clear increase in the prevalence and proportion of LLM-related terms after ChatGPT's release in late 2022. For instance, the terms delve[s/d/ing] and underscore[s/d/ing] had significant growth across different scholarly databases between 2022 and 2024 (1360% and 1062% in Scopus, respectively). In contrast, other common research terms such as investigate[s/d/ing] and highlight[s/ed/ing] had only slight increase (only 9% and 86%) over the same period. The term "underscore" appeared in 20% of PMC publications and 11% of Dimensions publications, indicating a considerable shift using it in academic writing. The 16% increase for underscore[s/d/ing] suggests that at least 16% of academic publications published in 2024 had their language influenced by LLMs, and the above discussion suggests that the overall figure for LLM influence is probably at least triple this (i.e., close to half).

In answer to the second research question, there were noticeable disciplinary differences in how LLM-related terms were used. For example, underscore[s/d/ing] was particularly prominent in Environmental Science (0.26% to 3.84%) and Business (0.38% to 3.54%).

In answer to the third research question, the use of LLM-related terms varied substantially between titles and abstracts. For instance, unveil[s/ed/ing] was more common in titles (0.04% to 0.26%), while underscore[s/d/ing] appeared more often in abstracts (0.21% to 2.53%) in 2022 and 2024 respectively.

Although this study provides new evidence that LLMs like ChatGPT may have influenced academic writing through the analysis of updated data, a broader range of terms, and multiple scholarly databases, further research is needed to understand how LLMs are shaping academic publishing across specific subjects, considering their relatively recent introduction. Different LLMs (e.g., ChatGPT, Gemini, and DeepSeek) may use unique terms when generating or editing academic texts. Hence, future research could investigate differences between LLMs in their influence on academic writing.

References

Elsevier. (2024). Insights 2024: Attitudes toward AI – Full report. Elsevier. https://www.elsevier.com/insights/attitudes-toward-ai

Eppler, M., Ganjavi, C., Ramacciotti, L. S., Piazza, P., Rodler, S., Checcucci, E., & Cacciamani, G. E. (2024). Awareness and use of ChatGPT and large language models: a prospective cross-sectional global survey in urology. European Urology, 85(2), 146-153.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digital Medicine, 6(1), 75.

Gray, A. (2024). ChatGPT" contamination": estimating the prevalence of LLMs in the scholarly literature. arXiv preprint arXiv:2403.16887.

Huang, Y., Wu, R., He, J., & Xiang, Y. (2024). Evaluating ChatGPT-4.0's data analytic proficiency in epidemiological studies: A comparative analysis with SAS, SPSS, and R. Journal of Global Health, 14.

Hwang, T., Aggarwal, N., Khan, P. Z., Roberts, T., Mahmood, A., Griffiths, M. M., ... & Khan, S. (2024). Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. Plos one, 19(2), e0297701.

Kacena, M. A., Plotkin, L. I., & Fehrenbacher, J. C. (2024). The use of artificial intelligence in writing scientific review articles. Current Osteoporosis Reports, 22(1), 115-121.

Khalifa, M., & Albadawy, M. (2024). Using artificial intelligence in academic writing and research: An essential productivity tool. Computer Methods and Programs in Biomedicine Update, 100145.

Kousha, K. (2024). How is ChatGPT acknowledged in academic publications? Scientometrics, 129(12), 7959-7969.

Lechien, J. R., Gorton, A., Robertson, J., & Vaira, L. A. (2024). Is ChatGPT-4 Accurate in Proofread a Manuscript in Otolaryngology–Head and Neck Surgery?. Otolaryngology–Head and Neck Surgery, 170(6), 1527-1530.

Liang, W., Zhang, Y., & Wu, Z. (2024). Mapping the increasing use of LLMs in scientific papers. arXiv. https://arxiv.org/pdf/2404.01268

Naddaf, M. (2025). How are researchers using AI? Survey reveals pros and cons for science. Nature. https://doi.org/10.1038/d41586-025-00343-5

Margetts, T. J., Karnik, S. J., Wang, H. S., Plotkin, L. I., Oblak, A. L., Fehrenbacher, J. C., ... & Movila, A. (2024). Use of AI language engine ChatGPT 4.0 to write a scientific review article examining the intersection of Alzheimer's disease and bone. Current Osteoporosis Reports, 22(1), 177-181.

Mishra, T., Sutanto, E., Rossanti, R., Pant, N., Ashraf, A., Raut, A., ... & Zeeshan, B. (2024). Use of large language models as artificial intelligence tools in academic research and publishing among global clinical researchers. Scientific Reports, 14(1), 31672.

Park, Y. J., Kaplan, D., Ren, Z., Hsu, C. W., Li, C., Xu, H., & Li, J. (2024). Can ChatGPT be used to generate scientific hypotheses?. Journal of Materiomics, 10(3), 578-584.

Picazo-Sanchez, P., & Ortiz-Martin, L. (2024). Analysing the impact of ChatGPT in research. Applied Intelligence, 54(5), 4172-4188.

Salvagno, M., Cassai, A. D., Zorzi, S., Zaccarelli, M., Pasetto, M., Sterchele, E. D., & Taccone, F. S. (2024). The state of artificial intelligence in medical research: A survey of corresponding authors from top medical journals. Plos one, 19(8), e0309208.

Uribe, S. E., & Maldupa, I. (2024). Estimating the use of ChatGPT in dental research publications. Journal of Dentistry, 149, 105275.

Van Noorden, R., & Perkel, J. M. (2023). AI and science: what 1,600 researchers think. Nature, 621(7980), 672-675.

Appendix

The Number of academic publications with LLM-associated terms

All 12 potentially LLM-associated terms increased significantly in academic publications from 2023, after ChatGPT was released in November 2022 (Figure 8). For example, in Dimensions, mentions of *delve* related terms ("delves," "delving," "delved,") increased from 30,329 in 2022 to 268,483 in 2024 (785% increase). Similarly, *underscore* related terms increased by 557%, and *showcase* related terms by 364%. In Scopus, mentions of *delve* in titles, abstracts, or keywords increased by 1,582% (from 1,852 in 2022 to 31,149 in 2024) with similar increases found for *underscore* (1,046%), *showcase* (397%), and *unveil* related terms (243%). In PubMed, *delve* and *underscore* increased by 1,491% and 688%, respectively. These trends suggest that LLMs like ChatGPT are increasingly being used in academic publications after about two years of its release.

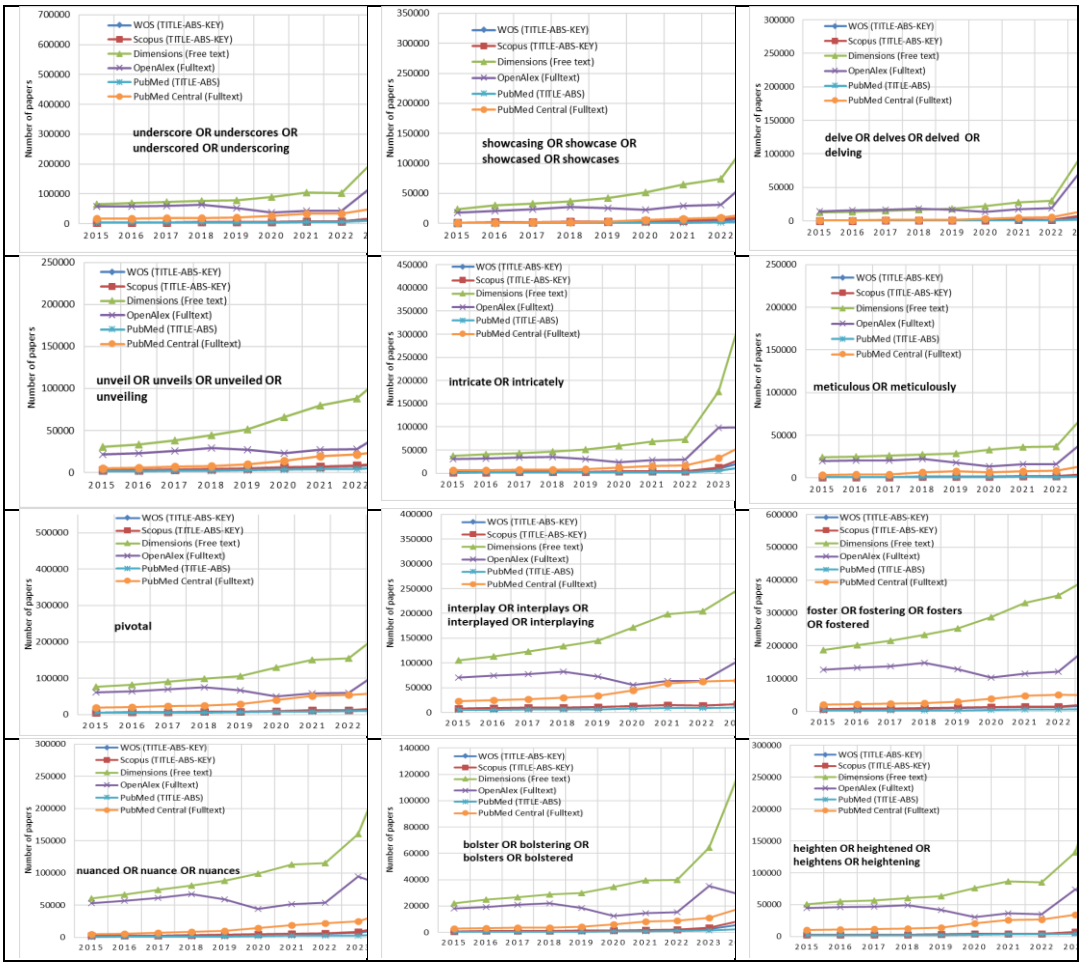


Figure A1. Number of academic publications (2015–2024) containing 12 potentially LLM-related terms across bibliographic and open-access databases.

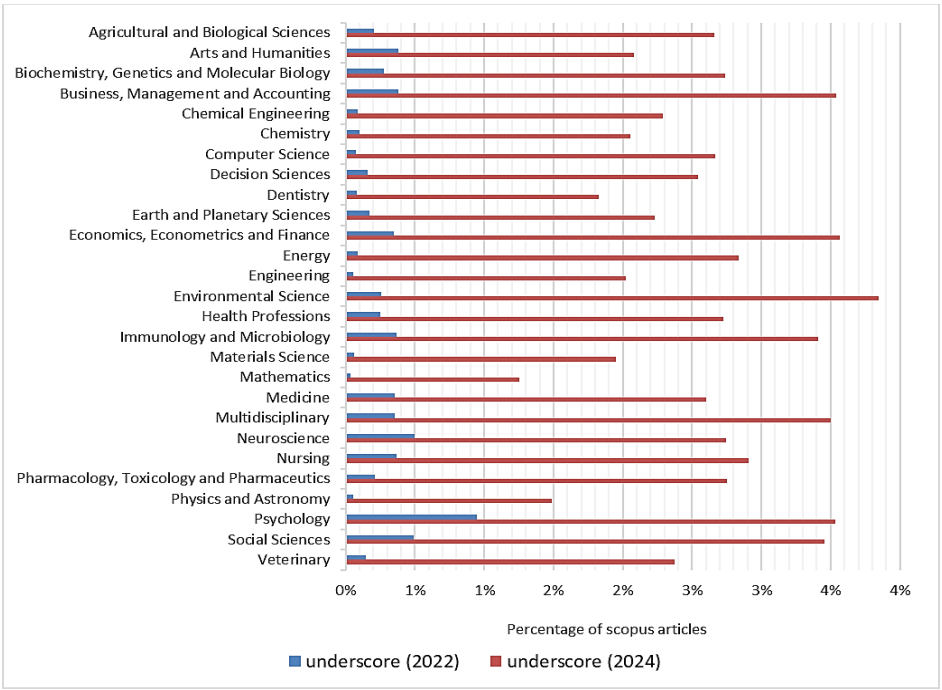


Figure A2. Percentage of Scopus papers with *underscore[s/d/ing]* in their title, abstract or keywords in 27 subjects (2022-2024).

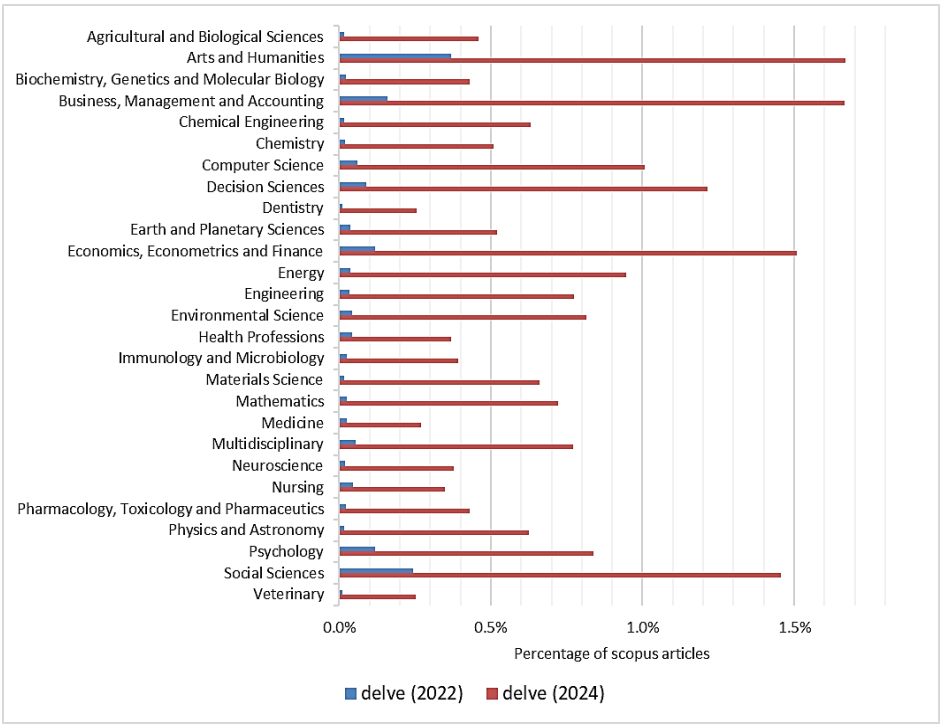


Figure A3. Percentage of Scopus papers with delve[s/d/ing] in their title, abstract or keywords in 27 subjects (2022-2024).

https://doi.org/10.51408/issi2025_055

How Scientific Research Impacts Policy Cycle

Ashraf Maleki[1], Marianna Lehtisaari[2], Kim Holmberg[3]

[1]*ashraf.maleki@utu.fi*, [2]*mkatal@utu.fi*, [3]*kim.j.holmberg@utu.fi*
Dept of Social Research, University of Turku, Assistentinkatu 7, FI-20014 Turku (Finland)

Abstract

Scientific research is increasingly referenced in policy documents issued by international, national, and regional organizations, reflecting its role in governance and decision-making across diverse social responsibilities. However, the extent to which scientific publications contribute to different stages of policy making remain an under-researched area. This study investigates how policy sources cite scientific research across disciplines, with a particular focus on the placement and function of the citations within governmental and intergovernmental organization (IGO) policy documents. Our core dataset is drawn from UK REF2021 journal articles, while policy citation counts and a sample of policy documents were retrieved from Overton.io. A random sample of 1,000 policy documents citing scientific articles in five fields from governmental and IGO sources were analyzed to determine type of policy documents, their purposes, and the placement of the citations in them. Policy documents, based on their focus and their purpose, were assessed according to the five-stage policy chain model: agenda-setting, formulation, adoption, implementation, and evaluation. The findings indicate that governmental and IGOs are the predominant sources of policy citations. Many policy documents lack distinct sections typical of scientific articles and appear in numbered chapters (41%), while in the remaining documents citations were primarily located in the Introduction (13%), Background (9%), Methods (10%), or References without clear in-text citation (7%). With some disciplinary differences, nearly half of policy citations appear in the "policy formulation" stage of the policy making chain, while about one-fifth occur in the "policy evaluation" stage, demonstrating how policymakers rely on academic research both when shaping policy frameworks and assessing their effectiveness. Field of public health stands out as an exception, with a significantly higher proportion of scientific citations in the "policy implementation" stage (34%) compared to other fields (8%), reflecting the evidence-based nature of practical guidance and guidelines informed by research. Additionally, most policy document sources had more administrative (63%) than scientific (37%) focus and held operational (39%), advisory (26%), or executive (16.5%) roles, highlighting their action-oriented nature. The results challenge the view that policy documents merely synthesize academic research; instead, they often engage in knowledge production through commissioned studies, empirical analysis (56%), and evidence-based recommendations (34%). Policy-to-research citations should not be seen solely as indicators of research uptake but as part of a reciprocal process where policy documents both utilize and generate scientific knowledge. Policy citations can thus serve as a critical measure of the impact of science on policy research and recommendations, demonstrating how academic research informs and shapes evidence-based governance.

Introduction

Policy citations, i.e. citations to scholarly research in so-called policy documents, is an understudied area of responsible use of metrics in research impact assessment. This paper investigates how scientific findings are cited in policy documents, their placement within the text, and their broader implications for policy. Overton.io identifies and aggregates citations to academic articles from policy documents that have been openly published online by various national or international, governmental or intergovernmental organisations, thus allowing large scale analysis of policy citations for research impact assessment. The policy citations collected by

Overton.io are mostly citations in grey literature hosted on governmental, inter-governmental and institutional websites. Although policy citations offer an opportunity to investigate citation context and societal impact of research in a more diverse way than many other altmetric indicators are able to, the diversity of governmental and intergovernmental activities and responsibilities calls for more research into the context and content of policy citations before they can be reliably used for research impact assessment. The current study investigates to what degree the analysed policy documents offer policy recommendations and how the citations to research have been used in them.

Governmental and intergovernmental organizations serve far-ranging purposes in policy development, from setting international frameworks to implementing national and regional regulations. These organizations create policy documents to establish priorities, provide evidence-based and strategic guidance, enforce regulatory standards, and disseminate best practices. By investigating the context in which research is being cited in policy documents we can deduce some new understanding of how scientific evidence informs policy decisions and how research informs different stages of the policy chain of formal decision-making.

The policy chain or policy cycle refers to the sequence of processes through which policy ideas are developed, implemented, and evaluated (Jann and Wegrich, 2017). The five-stage model of policy cycle consists of several key actions based on previous studies dating back to Jones (1974):

1. **Problem Identification and Agenda Setting.** The first stage involves recognizing and defining an issue that requires governmental or intergovernmental intervention, prioritizing policy issues for discussion, and determining which problems should receive attention from policymakers.
2. **Policy Formulation and Analysis**. The second stage involves designing potential solutions, strategies, or frameworks to address the identified issues and developing policy proposals, including drafting legislation, guidelines, and recommendations.
3. **Decision-Making and Policy Adoption.** The third stage involves selecting a specific course of action, which may involve legislative approval, executive orders, or administrative rulings and formalizing the decision.
4. **Policy Implementation.** Implementation involves operationalizing the selected policies and enforcing them through regulatory measures, public programs, or institutional actions.
5. **Policy Monitoring and Evaluation**. The final evaluative stage involves assessing the impact of policy measures and making necessary modifications based on empirical findings and stakeholder input.

Each of these steps is informed by scientific research in various ways and every stage benefits from other stages as policies keep evolving. By analyzing citations from policy documents to scientific research, this study aims to clarify how and to what extent scientific research informs different stages of policymaking and thus, what kind of impact research has had beyond academia.

Research Questions

The research aims to analyze policy citations to scientific publications and their placement both in the policy documents and in the policy chain. In order to address the research goal, the following research questions are addressed in this research:

1. How frequently is scientific research cited in policy documents across different fields and policy sources?
2. Where within policy documents do citations to scientific research typically appear, and how does their placement relate to policy development?
3. How do the characteristics of citing policy sources influence how scientific research is used at different policy chain stages?
4. What types of policy documents (research, policy recommendations, review, guidance/guideline, rules/regulations) cite scientific research, and how do these types correspond to different stages of the policy chain?

Background

Policy documents have emerged as a potential source for evidence of wider societal impact of research, i.e. how research is informing policy and through that, influencing society. Both Altmetric.com and Overton.io identify and collect online policy documents and extract and aggregate citations to scientific articles from them. While the societal impact of policy citations or their applicability for research assessment remain understudied areas of altmetric research, some studies have explored and compared these new data sources (Maleki and Holmberg, 2022; 2024; Murat et al., 2023; Dorta-González et al., 2024), pointing at some differences between them and suggesting that in order to gain a more robust picture data from both should be used. Earlier research has pointed at specific affordances that may have a positive influence on the likelihood for research to be cited in policy documents. It has been shown that research that has been discussed in blogs and news is more likely to also be cited in policy documents (Dorta-González et al., 2024). This suggests that science communicators may even have an important role in influencing policy. In line with the demand from many funders to approach complex societal challenges with multi-disciplinary research approaches, disciplinary diversity has been discovered to have a positive influence on the likelihood for research to be cited in policy documents (Pinheiro et al., 2021). Co-authorship with non-academic authors also appears to increase a research article's chances of getting cited in a policy document (Ma and Cheng, 2023).

There is some evidence that, at least in some research topics, policy citations identified and captured by Overton are associated with research impact, as measured by the peer-review assessment of impact by the UK Research Excellence Framework (REF) 2014 (Szomszor and Adie, 2022). The authors suggest that citation analyses on policy documents may be informative for research assessment and policy review. But contradictory results have also emerged. Research excellence, as measured with more traditional bibliometric measures, may not have an influence on whether a research article is cited in policy documents or not (Mahfouz et al., 2024). Based on these findings, it would appear that other, non-academic attributes and factors, may have a more important role in determining how research influences policy. More

research is needed before more definite conclusions about the applicability of policy citations for research assessment can be made.

The citation placements in policy documents have also been studied before. It has been discovered that citations to research appear most often in policy documents that could be classified as advice documents, and not in legislative or executive records (Pinheiro et al., 2021). When it comes to the context of the citations within the policy documents, one study showed that about half of the citations appeared in what the authors called "expounding" context (Yu et al., 2023). This included contexts that explained meaning or background, such as definitions of concepts, theoretical foundations, or argumentation of ideas. The context with the second largest amount (just over 20%) of citations was discovered to be sections that could be labelled as "review" sections.

While there are some earlier studies that have analysed various aspects of policy documents and policy citations, they still remain an understudied area of altmetric research. A greater understanding of policy citation patterns would allow for more meaningful evaluation of the applicability of policy citations for research impact assessment and inform about the role of research in policy making.

Data and Methodology

Our primary dataset of scientific publications constituted about 151,712 journal articles reported to REF2021. REF2021 refers to the overarching evaluation framework used to assess the quality and impact of research conducted in UK higher education institutions. These were examined for policy citations in Overton.io. In other words, policy documents indexed by Overton were searched for citations to scientific journal articles reported to REF2012. The policy citations were retrieved through Overton API during November 2024. The regular Unit of Assessments (UoAs) from REF2021 were used for subject classification of the journal articles and sampling of the data.

A sample of policy documents that were citing scientific publications were randomly selected to analyse the policy sources and their various purposes in the policy chain. About 80% of the scientific publications reported to REF2012 had at least one government or IGO policy citation, whereas these on average constitute about 70% of all policy citations (Figure 1). In our analysis we focused on policy documents by governmental or intergovernmental sources, as these constitute the majority of all policy documents. We randomly selected a sample of 1006 policy documents (about 500 government and 500 IGO) that cite journal articles in five REF2021 subject categories (about 100 from each field per source type). The included REF2021 UoAs are Public Health, Health Services and Primary Care, Engineering, Earth Systems and Environmental Sciences, Business and Management Studies, and Art and Design publications.

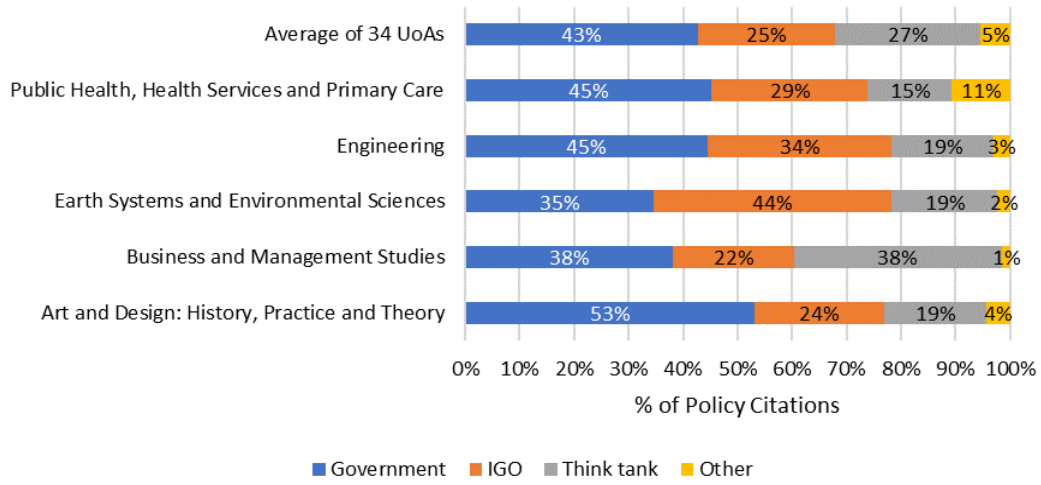


Figure 1. Cumulative Proportion of Policy Citations across source types to REF UoAs. (more in Appendix Table S2).

Appendix Table S1 gives the extent of the DOIs cited across the four Overton source types (government, IGO, think tank, and other) and appendix Table S2 gives the extent of citing policy document counts, both across the 34 UoAs.

The citation context, i.e., the sentence citing the scientific publication in the policy documents, and the placement of the citation in each document (e.g. introduction, method, findings, appendix etc.), were manually extracted from the policy documents in our sample.

While English is the majority language (851, 85%) of the sampled policy documents, the sample also contained documents in 23 other languages (155, 15%) (Table 1). The most frequent among these other languages were Spanish (47, 5%), French (30, 3%) and Swedish (18, 2%). Non-English documents were translated using Google Translate and purpose and citation contexts were extracted from the translated texts.

Our data sample had some limitations. A small proportion of the policy documents could not be reached online due to restricted access (6.3%) or because the document had been removed, while in a small number of documents the citation couldn't be located. Additionally, five documents had been withdrawn by linking to an updated or replacing document that contained the citations. Duplicate documents accounted for 1.7% of the sample. These were mostly versions of the same documents but in different versions or different drafts of the same manuscript. In some cases, some confusion was caused by the multiple PDF files that were associated with a single policy document.

Table 1. Characteristics of sampled policy documents (total n = 1006).

Characteristic	*Count*	*percent*
Language:		
English	851	84.6%
Non-English	155	15.4%
Lost sample:	66	6.6%
Online page could not be reached.	63	6.3%
Citation, wrong or not found	3	0.3%
Withdrawn	1	0.1%
Special cases:		
Withdrawn, but updated and replaced	5	0.5%
Duplicates (as different language and draft)	17	1.7%
Multi-PDF Policy documents	21	2.1%

To determine the purpose of the policy document the documents were manually searched for mentions of objectives or aims of the document. The identified texts containing a description of the purpose of the document were used to identify two aspects of documents: document type and policy chain stage. We identified document types based on possible presence of description of research, review, policy recommendation, guidance, or rules and regulations or some combinations of them. A significant effort was placed on identifying pure research from research that accompanies policy recommendation and between review studies and review studies with the main goal to advise policy (Table 2). Although policy recommendation is a potential outcome in most studies, not all studies have a similar emphasis on it and they may vary from purely academic research with potential policy advice to more focused policy advising research.

Table 2. Type of policy documents.

Type of policy documents	*Description*
Research	Original research in a journal or conference publications hosted on government websites. Do not necessarily offer policy recommendations.
Research & Policy Recommendation	Original research or analysis either in a scientific paper or organizational report with clear policy recommendations. Includes policy research papers.
Review	Systematic review/synthesis of original research and/or policies in a scientific paper or report meant for improving understanding of a subject or current literature. May contain a call for discussion or present various solutions, but do not offer any policy recommendation.
Review & Policy Recommendation	Systematic review/synthesis of original research and/or policies in a scientific paper or report meant for both improving understanding of a subject or current literature and offering a policy recommendation.
Guidance or Guideline	Guidance or guidelines, handbooks and evidence-based recommendations.
Rules and Regulations	Formal rules and regulatory documents.

Based on the identified purpose of the policy documents the documents were assigned to the five policy chain stages (Table 3). For this both manual coding and ChatGPT-4o were used (examples of texts used to identify different policy chain stages are in Appendix 1).

Table 3. Stages of the policy chain.

Policy chain stages	*Description*
1. Agenda-setting	Identifying issues that require government intervention.
2. Formulation	Developing possible policy solutions.
3. Adoption	Deciding which policies to implement.
4. Implementation	Putting policies into action.
5. Evaluation	Assessing the effectiveness of policies.

Because of the large number of sentences that were identified to contain evidence of either the purpose of the document or of its type, large language models (LLMs) were used for automated analysis of the textual content. An organizational premium access to ChatGPT was used to send a command prompt asking for ChatGPT to determine the policy document type and policy chain stages of the uploaded sentences containing policy document purposes. Several tests were conducted before the optimal command prompt was chosen.

To examine the reproducibility of the ChatGPT results, the prompts were repeated three times and any inconsistencies in the results were investigated. Documents for which the results changed were cross-checked manually by the first author to identify

potential reasons for changes in the classification results. Of the 1006 policy documents the classification of 81 documents changed between repeated tests with ChatGPT. ChatGPT was asked to assign the documents to only one of the given policy chain stages and to assign a policy document type to each document. The policy document types were manually coded by two humans and then using ChatGPT, while the policy chain stages were coded manually by one human and with the help of ChatGPT. Despite taking measures to try to secure reproducibility, it is unclear if the results can be reproduced if using ChatGPT. Updates in the LLM, users' previous interaction with the LLM, and certain level of disambiguation in the policy texts may influence the results at a later stage.

The sources of analysed policy documents were classified as governmental or intergovernmental; international, national or regional; administrative or scientific; and according to their function as advisory, research, executive, operational, legislative, regulatory, or information (Table 4).

Table 4. Policy source types with examples.

Source types	*Examples*
Governmental	*Government of Ireland (GOV.IE), County Administrative Boards (Sweden)*
Intergovernmental	Asian Development Bank, Arctic Council
International	*European Commission, International Monetary Fund (IMF)*
National	*Government of Singapore, NHS England*
Regional	*Government of Flanders, Public Health Wales*
Administrative	*European Parliament Committees, German Environment Agency (UBA)*
Scientific	*Joint Research Centre (European Commission), Eurostat*
Advisory	*Intergovernmental Science-Policy Platform on Biodiversity (IPBES), Internet Governance Forum, European Economic and Social Committee*
Research	*National Renewable Energy Laboratory (NREL), European Forest Institute (EFI)*
Executive	*New Zealand Treasury, Northern Ireland Executive*
Operational	*European Investment Bank, Food and Agriculture Organization of the United Nations, NHS Trusts*
Legislative	*European Parliament Plenary, Parliament of Denmark*
Regulatory	*Organisation for the Prohibition of Chemical Weapons, Bank of Italy*
Information	*Community Research and Development Information Service (CORDIS), Official State Gazette (Spain)*

Findings and Discussion

The findings show that of the 151,712 REF 2021 journal articles included in this study a total of 26% (n = 39,226) received a policy citation from the Overton data (Appendix Table S1), with 69% of them coming from government-sourced policy

documents, 46% from documents published by think tanks, 33% from documents by IGOs and 20% coming from other sources. The cumulative number of policy citations were 275,028 (Appendix Table S1) constituting 41% government policy citations, 28% IGOs, 27% think tanks and 6% other policy sources. This demonstrates that governmental documents are the most significant source of policy citations to journal articles. Although think tanks tend to cite a higher number of academic research (46% vs. 33% IGOs), IGOs produce almost similar extent of policy documents that are supported by research (28% vs. 27% think tank).

Policy Citations to REF2021 Articles

In answer to the first research question, Figure 2 indicates proportion of REF2021 journal articles with positive policy citations across the four Overton.io policy source types in 34 REF UoAs. In most fields, the government cites a significantly higher proportion of journal articles compared to other source types, with the highest coverage being in *Public Health, Health Services and Primary Care* (54%). The exceptions are *Economics and Econometrics*, and *Politics and International Studies* where think tanks (62% and 37%, respectively) cite a significantly higher proportion of scientific publications than government (42% vs. 20%).

Figure 3 indicates the geometric mean of policy citations across source types. While reflecting a similar trend as in the proportions of cited documents, the highest geometric mean policy citations were made by think tanks citing Economics and Econometrics journal articles (2.4), followed by government citations to P*ublic Health, Health Services and Primary Care* (1.3).

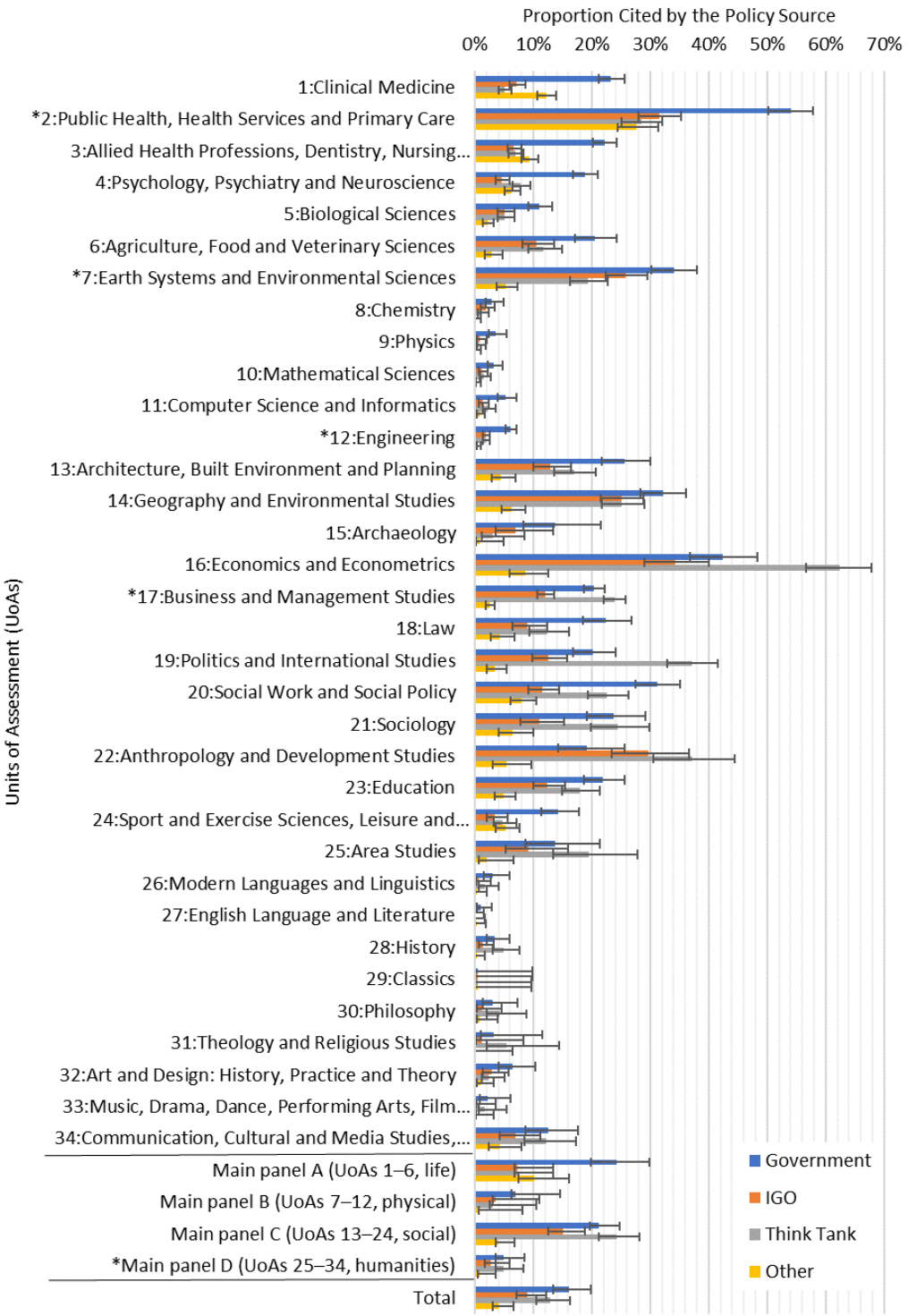


Figure 2. Proportion non-zero policy citations of REF journal articles (2014-2020) across UoAs across source types (government, IGO, think tank, and other). (Detailed in Appendix Table S1).

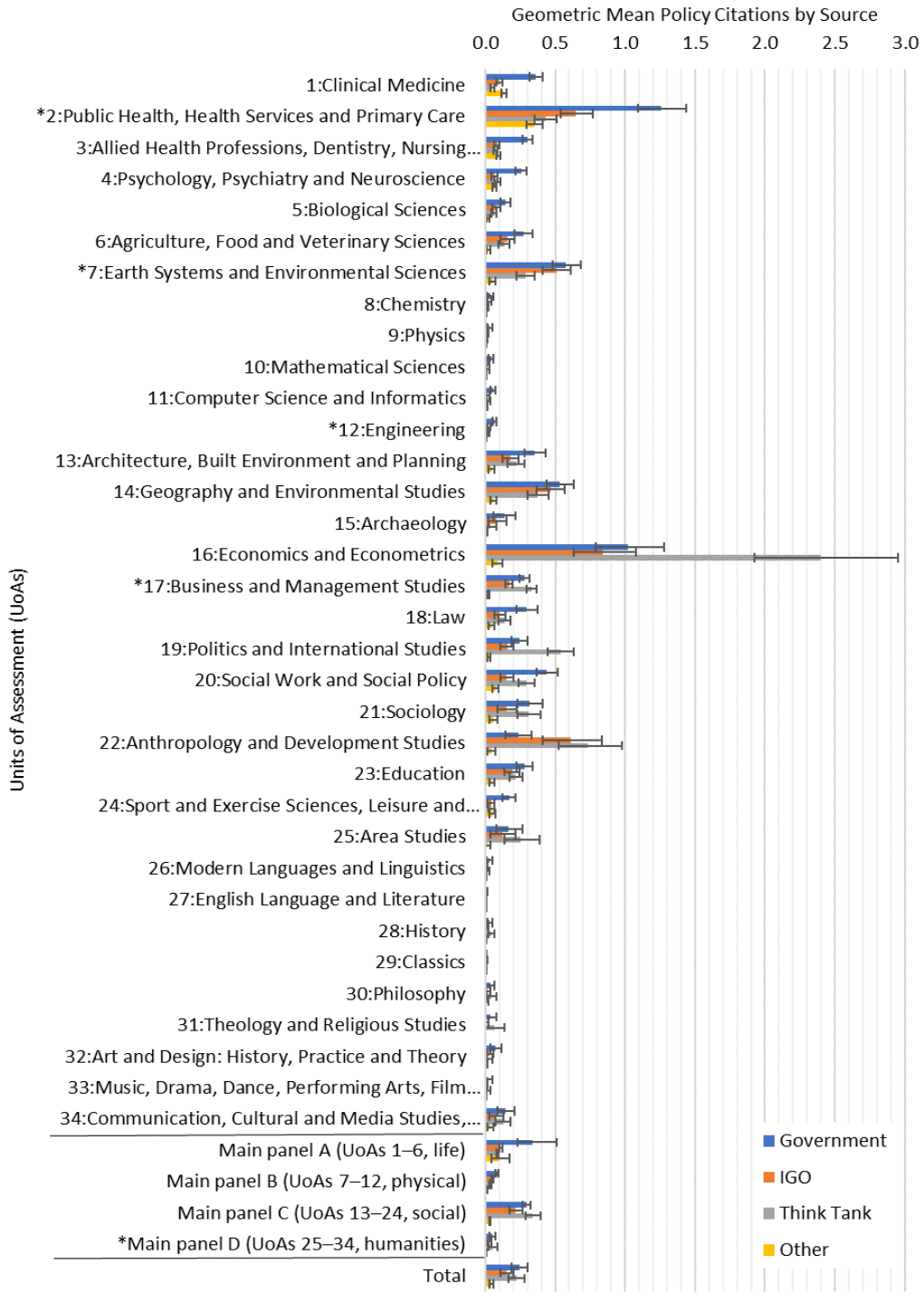


Figure 3. Average Geometric mean policy citations of REF 2021 journal articles (2014-2020) across UoAs in terms of citing source types (government, IGO, think tank, and other). * Fields used in the sampling.

Citation Placements in Policy Documents

In answer to the second research question, Figure 4 shows where in the analysed documents the citations to scientific articles were discovered. The policy documents often had several citations occurring in different parts of the document. Several policy documents in our sample contained numbered chapters rather than a structure that would have been similar to that of scientific journal articles (i.e. sections for executive summary, introduction, methodology, findings, discussion, conclusion, references and appendices). Thus, 41% of the citations were located simply in chapters, rather than some specific sections typical for a scientific article. Our categorization thus differs somewhat from that of Yu et al. (2023) in terms of citation locations in text. Of the citations that were found in specific sections most were discovered in Introduction (13%), Background (9%), Method (10%) and in References (7%). Figure 4 gives a breakdown of in which sections or parts of the policy documents the citations were found and the chapter numbers where the policy citations to research were found. Many of the policy documents had significantly more chapters than the ten reported here, but because the statistics dropped significantly after chapter 7 we decided to not report all the chapters. The results showed how the first chapters of the policy documents had most of the citations, resembling the structure of scientific articles where the earlier parts of the articles and reports alike, review and present earlier scientific evidence.

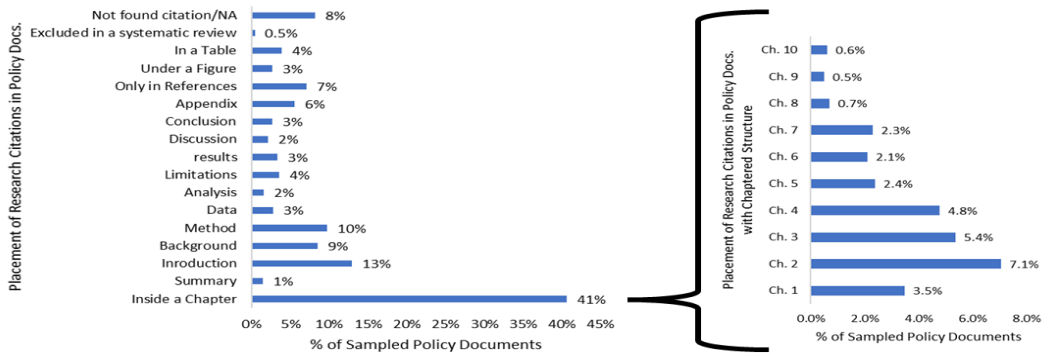


Figure 4. Proportion of sampled policy documents (total n = 1006) containing the REF2021 journal paper citations in terms of placement of citations in different headings of policy documents. The percentages are shown only for the first 10 chapters for brevity. Ch.: Chapter.

Characteristics of Policy Documents

Figure 5 gives a breakdown of specific characteristics of the policy documents. The proportion of documents from IGOs (66%) were almost double that of governmental documents (34%). There reasons for this contrast from the initial sample is that EU organizations have already been identified as governmental source types in Overton, contributing our sample from the government and making it sound larger than actually it was. But now that we applied our source type categorization, EU-level organizations migrated categorically to the Intergovernmental category, causing the

government sample to shrink. With an almost two thirds majority (65%) the policy documents came from international organizations, while only under a third came from national sources and regional sources contributed to only 7% of the documents. Majority of the policy citations came from organizations with administrative (63%) focus rather than scientific (37%), while their functions were mainly classified as operational (39%), advisory (36%) and executive (17%). Only 10% of policy sources held research functions, while only 5% were regulatory and 2% were legislative.

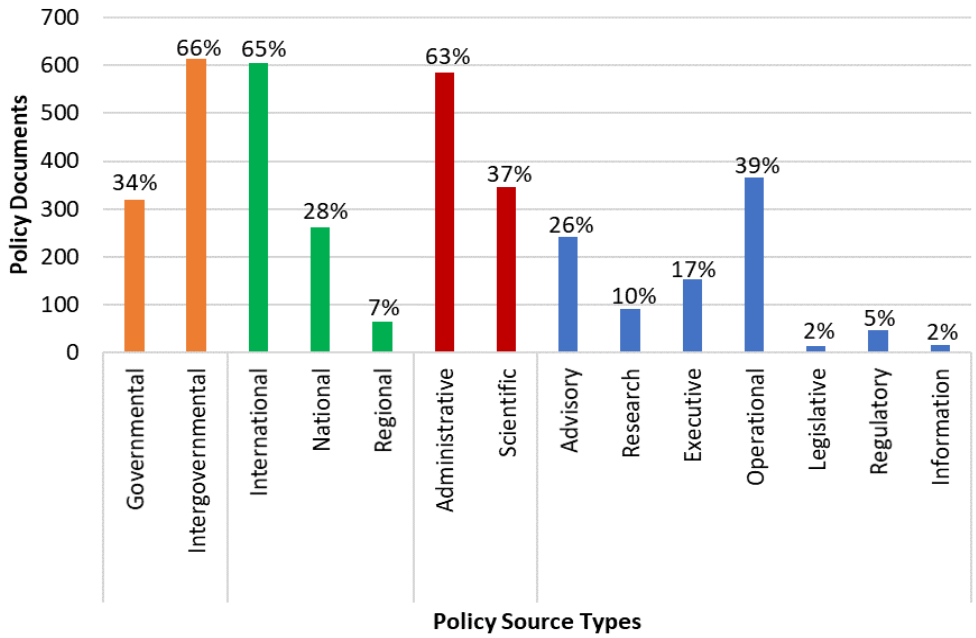


Figure 5. Count and proportion of sampled policy documents in terms of policy source type. (total n = 932).

Policy Source Functions and Policy Chain Stages

Table 5 shows how the functions of the policy documents align with the policy chain stages. The results demonstrate how almost half of the policy documents appear in the formulation stage of the policy chain and about one fifth appear in the evaluation stage, together accounting for 63% of policy citations (of total = 1006). These results reflect the importance of research in policy formulation on one hand and for the evaluation of the policy on the other hand. In other words, the results demonstrate that policy-makers often use academic research in the policy formulation and when developing indicators to assess effectiveness of policy approaches. It is also noteworthy that the majority of the policy documents appear to be operational (39%), advisory (26%) or executive (16.5%) in their function. This may reflect an action-oriented approach of the policy documents.

Table 5. Policy source structures mapped on policy chain stages of sample policy documents (total n = 932).

Source function	*1. Agenda-Setting*	*2.Formulation*	*3. Adoption*	*4. Implementation*	*5. Evaluation*	*Total*
Advisory	42 (4.5%)	**94 (10%)**	37 (4%)	18 (2%)	51 (5.5%)	**242 (26%)**
Research	9 (1%)	32 (3%)	4 (0.4%)	8 (0.9%)	**38 (4%)**	91 (10%)
Legislative	2 (0.2%)	6 (0.6%)	3 (0.3%)	1 (0.1%)	3 (0.3%)	15 (1.6%)
Regulatory	7 (0.8%)	**25 (3%)**	2 (0.2%)	1 (0.1%)	14 (1.5%)	49 (5%)
Executive	11 (1.2%)	**66 (7%)**	15 (2%)	24 (3%)	38 (4%)	**154 (16.5%)**
Operational	52 (6%)	**217 (23%)**	44 (5%)	22 (2%)	31 (3%)	**366 (39%)**
Information		**11 (1.2%)**			4 (0.4%)	15 (1.6%)
Total	123 (13%)	451 (48%)	105 (11%)	74 (8%)	179 (19%)	932 (100%)

Types of Policy Documents and Stages of the Policy Chain

The content of the citing policy documents is identified for policy document type analysis. One of the most important findings in this analysis is the identification of a high proportion of research-based policy documents (44%, 438), as they are hosted on government and intergovernmental organization (IGO) websites and cite journal articles. This is significant because policy documents are generally not associated with original research in the academic sense (i.e., primary research involving novel experiments, data collection, or theoretical development). However, our findings indicate that research is abundant as a type of policy document. This is likely because some organizations commission research studies specifically for policy-making, which can be considered gray literature rather than peer-reviewed original research. For instance, a World Bank report on economic development might include empirical analysis but is not peer-reviewed like journal articles. Additionally, our data suggest that the majority of policy sources conduct and publish original research as policy reports, which may contain surveys, statistical and cost-effectiveness analyses, or case studies. For example, some OECD (Organization for Economic Co-operation and Development) reports include novel data analysis on global education trends or World Bank reports published as research and policy brief quantifies the cost-effectiveness of various interventions to avert per death in a pandemic.

Policy documents are typically known for relying on and synthesizing existing research to inform policy recommendations. In our dataset, we found that Review & Policy Recommendation documents were the second most common type (23%, 232). Additionally, Research & Policy Recommendation (12%, 119) and Guidance and Guidelines (11%, 109) were also prevalent. These types of policy documents are often high-quality and evidence-based, meaning they integrate data and insights from academic research, stakeholder consultations, and case studies (e.g., health policy briefs and white papers). Only a minor proportion of policy documents (3%, 28) were identified as review papers only.

The policy document types were also examined against the policy chain stages (Figure 6). The results showed that *research* is the most common type of document, appearing mainly in policy formulation (21%), evaluation (13%) or agenda-setting (7%) policy chain stages. *Review and policy recommendation* is the second most common document type and they mainly contribute to policy formulation (10%) and policy adoption (6%). *Research and policy recommendation* are the third common document type, mainly appearing in the formulation (8%) stage of the policy chain. *Guidance and guidelines* are much less frequent overall but they make up a significant share of policy documents in the implementation stage (4%). Only a few *Rules and regulation* documents appeared in the policy adoption and the implementation stages (both 0.3%).

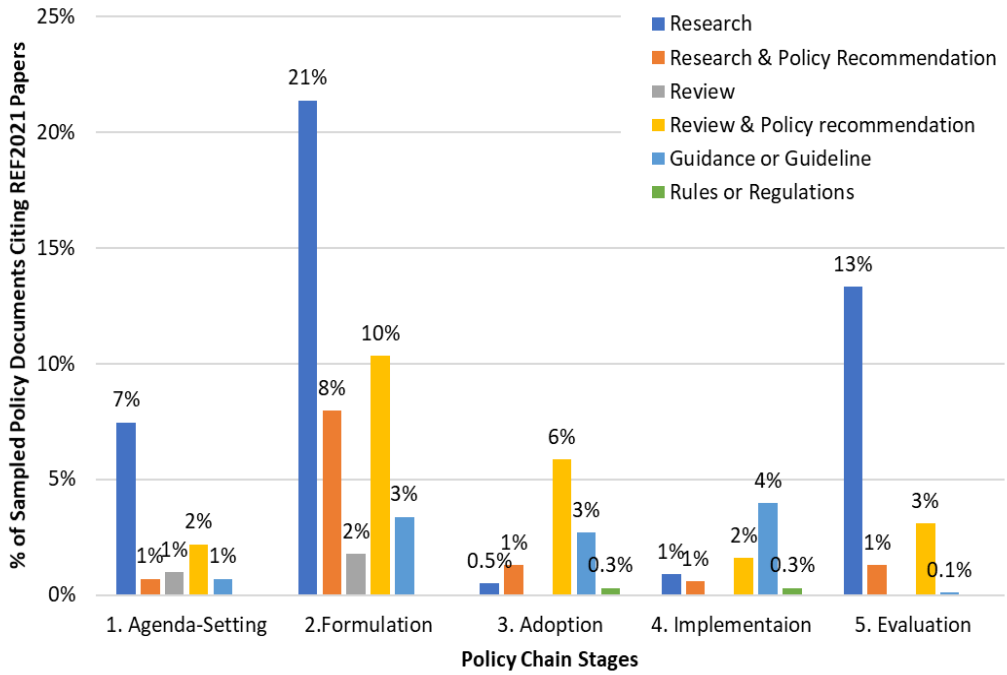


Figure 6. Proportion of all sampled policy documents (total n = 1006) across policy chain stages and policy document purposes.

Comparing Sampled Units of Assessment Across Policy Source Types and Policy Chain stages

Figure 7 shows that approximately 90% of all sampled policy documents could be traced for their role in the policy cycle. When comparing the cited research disciplines by the citing policy source types (governmental and intergovernmental) and policy chain stages, the results indicate that overall IGOs have a higher share of citing documents than governmental documents in both agenda-setting and in formulation stages, i.e. the early stages of the policy chain (Figure 7). The difference is the largest in Public Health, Health Services, and Primary Care, with governmental documents accounting for 5% of the agenda-setting documents, 17% of the formulation documents and a substantial 34% of implementation guidelines, while

policy documents by IGOs account for 7%, 46%, and only 10%, respectively. The results also demonstrate that depending on the UoAs between 5% and 18% of the policy documents were not accessible at the time of the analysis and thus, resulting in missing data in the analyses.

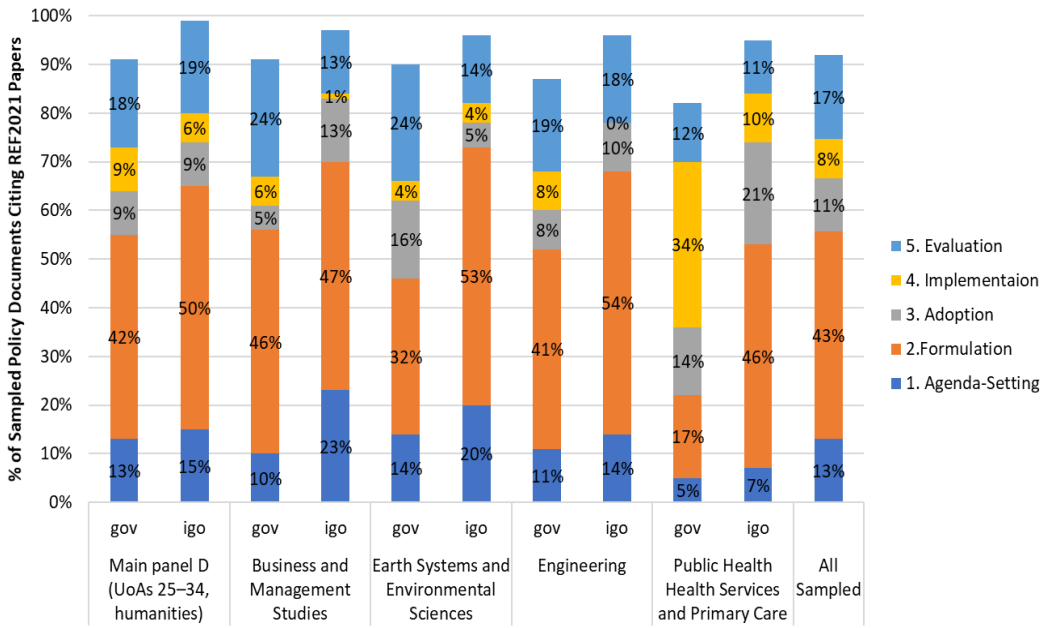


Figure 7. The proportions of policy documents by governmental and IGO sources across policy chain stages compared across five subject fields.

Conclusion

This research investigated the role of policy citations in understanding the broader societal impact of academic research. By analyzing how policy documents cite scientific publications, the results of this research offer insights into the placement of the citations within the policy documents, the purpose of the policy documents, and how they are placed within the policy-making process. Our findings confirm that governmental and intergovernmental organizations (IGOs) are the predominant sources of policy citations, with governmental policy documents containing the majority of the citations to policy-relevant research. Policy citations offer a promising data source for assessing research impact beyond traditional academic metrics, although the interpretation of the policy citations require careful consideration of the policy context in which they appear.

Our analysis reveals notable patterns in how citations appear within policy documents. Unlike scientific journal articles, which follow a standardized structure, many policy documents are organized into numbered chapters without distinct sections typical for scientific articles. As a result, 41% of the policy citations were found in chapters, while the remainder were primarily located in sections such as Introduction (13%), Background (9%), Methods (10%), and only in References without clear in-text citation (7%). In both cases, chapters and sections alike, policy

citations appeared mainly in the earlier parts of the policy documents, resembling the structure of academic papers where prior research is reviewed before new contributions are presented.

Our findings contribute to our understanding of the role research has in informing different stages of the policy making process. Almost half of the policy citations appear in the formulation stage of the policy chain, while about one-fifth are in the evaluation stage, suggesting that policymakers frequently rely on academic research when in the beginning shaping policy frameworks and when in the end assessing their effectiveness. It was also discovered that the source of most policy documents was operational (39%), advisory (26%), or executive (16.5%) functions, reflecting their action-oriented nature.

Policy-to-research citations should be interpreted with nuance, as policy documents are not merely consumers of existing research but, in many cases, also producers of original research. The high proportion of research-based policy documents (44%) suggests that policy documents often engage in empirical work, even if they are not traditionally considered as academic research. This challenges the simplistic interpretations of policy citations as one-way knowledge transfer and instead highlights a more dynamic and interactive relationship where policy documents contribute to, synthesize, and sometimes generate research. The prominence of Review & Policy Recommendation documents (23%) further indicates that policy documents often rely on academic work while framing it within broader policy contexts. Therefore, when analyzing policy citations, it is important to account for the diverse roles of policy documents - not just as passive users of research but as active participants in knowledge production and dissemination.

The findings from our research offer an original perspective that investigates policy documents through their placement in different policy chain stages and the role of policy citations in research impact assessment. The diversity in the placement of policy citations and the different document types points to the need to reconsider and to refine methodologies for interpreting policy citations. Future studies could explore qualitative aspects of citation use, assess disciplinary differences, and further investigate the mechanisms by which research informs policy decisions. By deepening our understanding of how policy citations function within the broader policy making landscape, we can improve their reliability as indicators of societal impact in research assessment frameworks.

Acknowledgments

The authors would like to express their gratitude to Dr. Kayvan Kousha (University of Wolverhampton) and Professor Mike Thelwall (University of Sheffield) for generously sharing the REF dataset and for their valuable contributions in shaping the initial development of the policy content analysis framework for this research. Their insights and support have been instrumental in advancing this study.

Funding information

This study is part of the research project Applicability of altmetrics in research impact assessment, funded by the Academy of Finland (332961).

References

Dorta-González, P., Rodríguez-Caro, A., & Dorta-González, M. I. (2024). Societal and scientific impact of policy research: A large-scale empirical study of some explanatory factors using Altmetric and Overton. Journal of Informetrics, 18(3), 101530. https://doi.org/10.1016/j.joi.2024.101530

Jones, C. O. (1974). Doing before knowing: concept development in political research. *American Journal of Political Science*, *18*(1), 215-228. https://www.jstor.org/stable/pdf/2110663.pdf

Jann, W., & Wegrich, K. (2017). Theories of the policy cycle. In *Handbook of public policy analysis* (pp. 69-88). Routledge.

Ma, J., & Cheng, Y. (2024). Why do some academic articles receive more citations from policy communities?. *Public Administration Review*. 1-23. https://doi.org/10.1111/puar.13857

Mahfouz, B., Capra, L., & Mulgan, G. (2024). Assessing the influence of research quality on policy citations: Quantitative analysis finds non-academic factors more likely to influence how papers get cited in SDG policy. *Sustainable Development*. https://doi.org/10.1002/sd.3214

Maleki, A., & Holmberg, K. (2022). Comparing coverage of policy citations to scientific publications in Overton and Altmetric. com: Case study of Finnish research organizations in Social Science. *Informaatiotutkimus*, *41*(2–3), 92-96. https://doi.org/10.23978/inf.122592

Maleki, A., & Holmberg, K. (2024). Policy Citations tracked by Overton. io versus Altmetric. com: Case Study of Finnish Research Organizations in Social Sciences. *Informaatiotutkimus*, *43*(3-4), 4-28. https://doi.org/10.23978/inf.145570

Murat, B., Noyons, E., & Costas, R. (2023, April). Exploratory analysis of policy document sources in Altmetric. com and Overton. In 27th International Conference on Science, Technology and Innovation Indicators (STI 2023). International Conference on Science, Technology and Innovation Indicators.

Pinheiro, H., Vignola-Gagné, E., & Campbell, D. (2021). A large-scale validation of the relationship between cross-disciplinary research and its uptake in policy-related documents, using the novel Overton altmetrics database. Quantitative Science Studies, 2(2), 616-642. https://doi.org/10.1162/qss_a_00137

Szomszor, M., & Adie, E. (2022). Overton: A bibliometric database of policy document citations. *Quantitative science studies*, *3*(3), 624-650. https://doi.org/10.1162/qss_a_00204

Tattersall, A., & Carroll, C. (2018). What can Altmetric. com tell us about policy citations of research? An analysis of Altmetric. com data for research articles from the University of Sheffield. *Frontiers in research metrics and analytics*, *2*, 9. https://doi.org/10.3389/frma.2017.00009

Yu, H., Murat, B., Li, J., & Li, L. (2023). How can policy document mentions to scholarly papers be interpreted? An analysis of the underlying mentioning process. Scientometrics, 128(11), 6247-6266. https://doi.org/10.1007/s11192-023-04826-y

Appendix

Examples of texts used to identify different policy chain stages:

1. Agenda-setting and Policy Problem Identification.

Example 1:

"The report presents scientific and technical background intended to stimulate debate and serves as a basis for further work to achieve a harmonized European view on the design and verification of such structures."

Example 2:

"This consultation paper is an initial consultation that sets out Central Bank proposals and seeks views on the introduction of a tiered regulatory approach for credit unions."

2. Policy Formulation and Analysis.

Example 1:

"This paper therefore provides empirical evidence in support of theoretical work stressing the importance of domestic variables in determining sudden stop episodes complementing the recent empirical literature which found a predominant role for global factors."

Example 2:

"The work presented in this report attempts to explore other realms about the future(s) of work beyond the strongly driven narrative of digital transformation. We have addressed one particular grassroots community, the Maker Movement, which is de facto enabling new models of education, collaborative work, and manufacture."

3. Policy Adoption and Decision-Making.

Example 1:

"In our discussions we shared five main goals: • secure and guarantee the necessary extra investment; • make practical changes to help solve the big challenges facing general practice, not least workforce and workload; • deliver the expansion in services and improvements in care quality and outcomes set out in The NHS Long Term Plan, phased over a realistic timeframe; • ensure and show value for money for taxpayers and the rest of the NHS, bearing in mind the scale of investment; • get better at developing, testing and costing future potential changes before rolling them out nationwide."

Example 2:

"This paper provides decision-makers with a framework for prioritising different economic, social and environmental goals and analysing the options available to achieve them. To this end, it develops three stylised COVID-19 recovery pathways ("Rebound", "Decoupling" and "Wider well-being") that differ in the extent to which they encompass greenhouse gas (GHG) emission reductions and the integration of mitigation and wider well-being outcomes or, broadly equivalently, SDGs."

4. Policy Implementation.

Example 1:

"This guideline covers identifying and managing familial hypercholesterolaemia (FH), a specific type of high cholesterol that runs in the family, in children, young people and adults. It aims to help identify people at increased risk of coronary heart disease as a result of having FH."

Example 2:

"This handbook aims at helping its users to effectively co-create the powerful policies we need today. It combines an entrepreneurial way of thinking and a concrete process for developing breakthrough ideas that stand a high chance of producing real-world impact. It presents a practitioner-oriented narrative for the design and implementation of innovative participatory processes and workshops to address societal challenges – coordinated by policymakers and with the active engagement of key stakeholders."

5. Policy Evaluation and Monitoring.

Example 1:

"To evaluate effectiveness and harms of opioids compared to nonopioid analgesics as treatment of moderate to severe acute pain in the prehospital setting."

Example 2:

"The Assessment Report on Land Degradation and Restoration by the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (IPBES) provides a critical analysis of the state of knowledge regarding the importance, drivers, status, and trends of terrestrial ecosystems."

Appendix Table S1. Frequency and percentage of journal articles in REF2021 with Overton.io policy citations across policy source types.

REF2021 Subjects	*DOIs*	*Policy Cited*	*by government*	*by IGO*	*by think tank*	*by other*
Agriculture, Food and Veterinary Sciences	3421	1022 (30%)	706 (69%)	368 (36%)	407 (40%)	100 (10%)
Allied Health Professions, Dentistry, Nursing and Pharmacy	11,547	3372 (29%)	2649 (79%)	800 (24%)	830 (25%)	1145 (34%)
Anthropology and Development Studies	1270	599 (47%)	236 (39%)	372 (62%)	465 (78%)	66 (11%)
Archaeology	790	152 (19%)	112 (74%)	53 (35%)	23 (15%)	6 (4%)
Architecture, Built Environment and Planning	3060	1100 (36%)	792 (72%)	397 (36%)	511 (46%)	139 (13%)
Area Studies	818	231 (28%)	112 (48%)	74 (32%)	162 (70%)	17 (7%)
Art and Design: History, Practice and Theory	1764	165 (9%)	115 (70%)	51 (31%)	41 (25%)	17 (10%)
Biological Sciences	7097	1096 (15%)	808 (74%)	376 (34%)	390 (36%)	162 (15%)
*Business and Management Studies	15,488	5708 (37%)	3200 (56%)	1866 (33%)	3735 (65%)	385 (7%)
Chemistry	3688	155 (4%)	114 (74%)	64 (41%)	36 (23%)	4 (3%)
Classics	244	7 (3%)	4 (57%)	2 (29%)	1 (14%)	1 (14%)
Clinical Medicine	11,971	3823 (32%)	3031 (79%)	913 (24%)	664 (17%)	1676 (44%)
Communication, Cultural and Media Studies, Library and Information Management	1542	369 (24%)	190 (51%)	112 (30%)	190 (51%)	69 (19%)
Computer Science and Informatics	5510	391 (7%)	294 (75%)	73 (19%)	116 (30%)	45 (12%)
*Earth Systems and Environmental Sciences	4365	1913 (44%)	1545 (81%)	1183 (62%)	901 (47%)	244 (13%)
Economics and Econometrics	2121	1473 (69%)	913 (62%)	737 (50%)	1339 (91%)	190 (13%)
Education	4133	1475 (36%)	934 (63%)	519 (35%)	751 (51%)	199 (13%)
*Engineering	17,963	1422 (8%)	1144 (80%)	350 (25%)	361 (25%)	100 (7%)
English Language and Literature	1962	31 (2%)	20 (65%)	2 (6%)	4 (13%)	7 (23%)
Geography and Environmental Studies	4162	1951 (47%)	1367 (70%)	1071 (55%)	1055 (54%)	275 (14%)
History	2633	210 (8%)	93 (44%)	41 (20%)	135 (64%)	10 (5%)
Law	2817	904 (32%)	620 (69%)	247 (27%)	338 (37%)	123 (14%)
Mathematical Sciences	5783	253 (4%)	197 (78%)	65 (26%)	88 (35%)	17 (7%)
Modern Languages and Linguistics	1821	83 (5%)	57 (69%)	13 (16%)	30 (36%)	6 (7%)

Music, Drama, Dance, Performing Arts, Film and Screen Studies	1071	40 (4%)	26 (65%)	6 (15%)	19 (48%)	4 (10%)
Philosophy	1126	69 (6%)	36 (52%)	13 (19%)	48 (70%)	11 (16%)
Physics	5480	280 (5%)	248 (89%)	39 (14%)	38 (14%)	10 (4%)
Politics and International Studies	3509	1668 (48%)	719 (43%)	445 (27%)	1331 (80%)	122 (7%)
Psychology, Psychiatry and Neuroscience	9718	2425 (25%)	1925 (79%)	476 (20%)	800 (33%)	679 (28%)
*Public Health, Health Services and Primary Care	4898	3553 (73%)	2765 (78%)	1591 (45%)	1444 (41%)	1472 (41%)
Social Work and Social Policy	4102	1843 (45%)	1286 (70%)	476 (26%)	930 (50%)	333 (18%)
Sociology	1975	783 (40%)	472 (60%)	218 (28%)	489 (62%)	126 (16%)
Sport and Exercise Sciences, Leisure and Tourism	3447	630 (18%)	467 (74%)	116 (18%)	158 (25%)	174 (28%)
Theology and Religious Studies	416	30 (7%)	13 (43%)	5 (17%)	21 (70%)	(0%)
Grand Total	151,712	39,226 (26%)	27,210 (69%)	13134 (33%)	17,851 (46%)	7934 (20%)

Appendix Table S2. Frequency and percentage of Overton.io-indexed policy documents citing REF2021 journal articles across policy source types.

REF2021 Subjects	*Policy Citations*	*Government*	*IGO*	*Think tank*	*Other*
Agriculture, Food and Veterinary Sciences	5351	2372 (44%)	1806 (34%)	1038 (19%)	148 (3%)
Allied Health Professions, Dentistry, Nursing and Pharmacy	16,791	9343 (56%)	3303 (20%)	1987 (12%)	2186 (13%)
Anthropology and Development Studies	5942	669 (11%)	2627 (44%)	2562 (43%)	84 (1%)
Archaeology	412	206 (50%)	144 (35%)	56 (14%)	6 (1%)
Architecture, Built Environment and Planning	6664	2537 (38%)	2000 (30%)	1925 (29%)	251 (4%)
Area Studies	1051	291 (28%)	254 (24%)	483 (46%)	23 (2%)
Art and Design: History, Practice and Theory	538	286 (53%)	128 (24%)	101 (19%)	23 (4%)
Biological Sciences	8253	3686 (45%)	2726 (33%)	1609 (19%)	263 (3%)
*Business and Management Studies	34,000	13,054 (38%)	7547 (22%)	13,061 (38%)	510 (2%)
Chemistry	725	366 (50%)	252 (35%)	102 (14%)	4 (1%)
Classics	10	4 (40%)	4 (40%)	1 (10%)	1 (10%)
Clinical Medicine	24,380	14,560 (60%)	5039 (21%)	1907 (8%)	3827 (16%)
Communication, Cultural and Media Studies, Library and Information Management	1279	462 (36%)	351 (27%)	376 (29%)	92 (7%)

Computer Science and Informatics	1417	741 (52%)	251 (18%)	348 (25%)	77 (5%)
*Earth Systems and Environmental Sciences	21,311	7507 (35%)	9484 (45%)	4223 (20%)	489 (2%)
Economics and Econometrics	32,567	7700 (24%)	7516 (23%)	17,494 (54%)	333 (1%)
Education	7877	2662 (34%)	2690 (34%)	2237 (28%)	336 (4%)
*Engineering	7046	3620 (51%)	2731 (39%)	1532 (22%)	243 (3%)
English Language and Literature	43	26 (60%)	4 (9%)	4 (9%)	9 (21%)
Geography and Environmental Studies	19,674	6195 (31%)	8578 (44%)	4703 (24%)	463 (2%)
History	477	147 (31%)	80 (17%)	239 (50%)	11 (2%)
Law	3395	1808 (53%)	623 (18%)	759 (22%)	240 (7%)
Mathematical Sciences	1574	865 (55%)	422 (27%)	360 (23%)	53 (3%)
Modern Languages and Linguistics	158	94 (59%)	19 (12%)	40 (25%)	6 (4%)
Music, Drama, Dance, Performing Arts, Film and Screen Studies	96	47 (49%)	13 (14%)	29 (30%)	7 (7%)
Philosophy	308	100 (32%)	52 (17%)	129 (42%)	27 (9%)
Physics	579	378 (65%)	115 (20%)	75 (13%)	11 (2%)
Politics and International Studies	7348	1798 (24%)	1366 (19%)	4057 (55%)	147 (2%)
Psychology, Psychiatry and Neuroscience	12,860	7412 (58%)	2134 (17%)	2407 (19%)	1327 (10%)
*Public Health, Health Services and Primary Care	37,375	17,689 (47%)	11,215 (30%)	6018 (16%)	4165 (11%)
Social Work and Social Policy	8843	3976 (45%)	2014 (23%)	2796 (32%)	547 (6%)
Sociology	4174	1446 (35%)	1175 (28%)	1343 (32%)	210 (5%)
Sport and Exercise Sciences, Leisure and Tourism	2428	1266 (52%)	561 (23%)	312 (13%)	289 (12%)
Theology and Religious Studies	82	26 (32%)	6 (7%)	50 (61%)	(0%)
Grand Total	275,028	113,339 (41%)	77,230 (28%)	74,363 (27%)	16,408 (6%)

https://doi.org/10.51408/issi2025_077

Identifying Vibrant Actors in Technology Development Through Their R&D Activity and Persistence

You-Fu Lee[1], Dar-Zen Chen[2], Chun-Chieh Wang[3]

[1]*d10522022@ntu.edu.tw*
National Taiwan University, Department of Mechanical Engineering, Taiwan (R.O.C.)

[2]*dzchen@ntu.edu.tw*
National Taiwan University, Department of Mechanical Engineering, Taiwan (R.O.C.)
Institute of Industrial Engineering, Taiwan (R.O.C.)

[3]*wangcc@ntu.edu.tw*
National Taiwan University, Department of Bio-Industry Communication and Development, Taiwan (R.O.C.)
Center for Research in Econometric Theory and Applications (CRETA), Taiwan (R.O.C.)

Abstract

Identifying vibrant actors in technological development is crucial for understanding innovation ecosystems and driving sustained advancements across industries. While traditional methods for identifying key contributors often focus on quantitative metrics such as patent counts and citation frequencies, they may overlook the persistence of R&D efforts—a critical factor in evaluating long-term technological impact. This study proposes a novel framework that incorporates both activity and continuity indicators to assess the sustained contributions of key actors in technology development. By applying a sliding window approach over a three-year period, this framework enables the identification of vibrant assignees who demonstrate consistent and impactful R&D engagement. The empirical analysis focuses on solid-state electrolyte technology for lithium batteries, a rapidly evolving field crucial to energy storage innovations. The study analyzed patent data from the United States Patent and Trademark Office (USPTO) from 2002 to 2021, identifying 981 relevant patents attributed to 223 assignees. The results reveal that while some assignees exhibit high patent counts, only a subset demonstrate persistent innovation over time, as captured through the proposed continuity index. Vibrant assignees, such as Samsung Electronics and LG Energy Solution, maintain consistently high continuity values, highlighting their strategic commitment to technological progress. In contrast, several non-vibrant assignees, despite holding substantial patent portfolios, lack sustained contributions, emphasizing the need to consider persistence in addition to patent volume when evaluating influence within an innovation ecosystem. The study's findings have significant implications for policymakers, industry stakeholders, and academic institutions, offering a more comprehensive approach to tracking and fostering technological leadership. Moreover, the proposed framework can be extended to various industries beyond energy storage, such as artificial intelligence and biotechnology, to analyze vibrant actors across different technological domains. Additionally, future research can apply this methodology to academic research institutions by analyzing journal publication data to evaluate the sustained contributions of universities and research organizations. Furthermore, the approach can be refined to assess individual inventors and authors, providing insights into their long-term impact and influence in their respective fields. In conclusion, this study advances the understanding of technological development by emphasizing the importance of persistence in R&D efforts. The proposed framework offers a robust tool for identifying vibrant actors, enabling more strategic resource allocation and fostering sustainable innovation in both industrial and academic settings.

Introduction

Understanding the key actors in technological development and R&D processes across various industries is crucial for deciphering the dynamics of innovation ecosystems. These actors serve as pivotal drivers of technological advancements, shaping industry trajectories and contributing to economic growth. Accurately identifying such contributors provides essential insights that inform policy decisions, guide investment strategies, and foster strategic collaborations among stakeholders. The identification of these key contributors is not only pertinent to academic research but also has practical implications for business strategies, government policies, and industry innovation planning (Valkokari et al., 2016).

Existing Approaches to Identifying Key Actors

A variety of methodologies have been employed to identify key actors within technology development processes, with patentometrics emerging as a particularly prominent approach. Valkokari et al. (2016) employed a design science framework to analyze the interactions between actors, resources, and activities within innovation ecosystems. Their study underscored the multifaceted roles that stakeholders play in driving innovation and highlighted the importance of coordinated efforts among diverse entities.

Dolphin and Pollitt (2020) advanced this field by applying machine learning techniques to UK patent data, enabling the identification of innovative entities within the electricity supply industry. Additionally, numerous studies have leveraged metrics such as patent citations, co-patenting networks, and technological classifications to delineate the ecosystem of influential R&D performers (Cohen, Fernandes, & Godinho, 2024). These approaches have significantly enhanced our understanding of the structural and collaborative dimensions of innovation ecosystems.

While these methods have provided valuable insights, they often rely heavily on quantitative indicators such as patent counts and citation frequencies, which capture only a snapshot of innovative activity. Such methods may overlook the critical element of persistence—an actor's sustained contributions over time—which is essential for assessing their true influence and long-term role in technological development.

Research Gap: The Importance of Persistence in R&D

Despite the advances in identifying key actors, a significant gap remains in the current methodologies: the insufficient emphasis on the persistence of R&D output. Innovation is not solely characterized by sporadic contributions or singular breakthroughs; rather, it requires continuous effort, adaptability, and sustained impact over time. Many traditional approaches fail to consider this longitudinal dimension, which is crucial for recognizing vibrant actors who actively and persistently shape technological landscapes (Cohen, Fernandes, & Godinho, 2024).

For example, reliance on patent counts may undervalue actors who produce fewer patents but contribute disproportionately to breakthrough innovations or foundational technologies (Griliches, 1990). Similarly, citation-based metrics, while indicative of influence, may not capture the durability and continuity of an actor's contributions (Narin, Noma, & Perry, 1987; Fleming & Sorenson, 2004). Without incorporating persistence into the analysis, existing methods risk overlooking key players who are instrumental in sustaining technological progress over extended periods (Breschi, Malerba, & Orsenigo, 2000).

Objectives and Contribution of This Study

This study aims to address the identified gap by incorporating the persistence of R&D output as a critical factor in identifying vibrant actors in technological development. By evaluating not only the quantity and immediate impact of R&D contributions but also their consistency over time, we seek to establish a more comprehensive framework for assessing influence within innovation ecosystems (Cohen, Fernandes, & Godinho, 2024).

Our approach integrates traditional patentometric methods with novel metrics designed to capture the longitudinal dimension of R&D activity. This combination allows for a more nuanced understanding of innovation ecosystems, identifying actors who consistently contribute to technological advancements and are likely to continue driving innovation in the future (Narin, Noma, & Perry, 1987).

In doing so, this study offers both theoretical and practical contributions. Theoretically, it enriches the literature on innovation ecosystems by highlighting the importance of persistence as a determinant of influence (Griliches, 1990; Fleming & Sorenson, 2004). Practically, it provides policymakers, industry leaders, and academic researchers with a robust tool for identifying key contributors, enabling more informed decisions regarding resource allocation, collaboration opportunities, and strategic investments (Breschi, Malerba, & Orsenigo, 2000).

Literature Review

Technology Development Dominated by Few Actors

Technology development across various industries is often driven by a limited number of key actors who play a crucial role in advancing innovation and shaping industry trends. Studies have shown that a small number of firms hold a significant share of patents in specific technological sectors, underscoring their pivotal influence on technological progress (Cohen, Fernandes, & Godinho, 2024). Notably, multinational corporations such as IBM, Samsung, and Siemens are frequently cited as leading innovators in their respective fields. The concentration of technological expertise within these dominant players highlights the importance of accurately identifying and analyzing their contributions to better understand the dynamics of innovation ecosystems (Valkokari, Amitrano, Bifulco, & Valjakka, 2016).

The dominance of a few key actors has significant implications for industry structure and competition. Breschi, Malerba, and Orsenigo (2000) found that a handful of firms control the majority of patents in the biotechnology and pharmaceutical

sectors, exerting substantial influence on the direction of technological change. This concentration of power can create high entry barriers for new entrants, potentially stifling competition and leading to monopolistic market conditions. Smaller firms and startups may face challenges in accessing critical technologies, limiting their ability to innovate and compete effectively.

Moreover, these dominant actors often have the capacity to influence standard-setting processes, regulatory policies, and industry norms, further solidifying their critical role in technological development (Blind, 2012). Their control over intellectual property can result in significant negotiation power, influencing licensing agreements and collaborative ventures. As a result, understanding the long-term influence of these key actors is essential for policymakers aiming to create balanced and inclusive innovation policies.

In addition, dominant players in technological development tend to form alliances and strategic partnerships that further strengthen their positions. Such collaborations enable resource-sharing and risk mitigation but can also result in knowledge silos, where technological advances remain confined to a select group of companies, limiting the broader diffusion of innovation. Therefore, examining how these firms sustain their dominance and identifying emerging challengers are crucial aspects of understanding the evolving innovation landscape.

Patentometrics for Identifying Key R&D Actors

Patentometrics has emerged as a powerful tool for identifying key R&D actors by utilizing quantitative measures derived from patent data to evaluate the innovation activities and impact of various entities. Valkokari et al. (2016) emphasized the importance of managing actors, resources, and activities within innovation ecosystems using a design science approach. Dolphin and Pollitt (2020) advanced this field by applying machine learning techniques to UK patent data, successfully identifying innovative entities within the electricity supply industry. Similarly, Hall, Jaffe, and Trajtenberg (2002) demonstrated the utility of patent citations as indicators of technological significance and innovation impact.

Further research has expanded on these methodologies to provide more nuanced insights. For instance, Huang, Notten, and Rasters (2011) employed network analysis to map co-patenting activities, revealing the collaborative networks that drive technological development. Such analyses help identify central actors who play key roles in fostering innovation and pinpoint potential areas for intervention to encourage broader participation in innovation ecosystems.

Additionally, Lerner and Seru (2017) explored patent text analysis to uncover the thematic focus of R&D activities, offering a deeper understanding of specific technological areas under development. Patent data, when analyzed in conjunction with text-mining techniques, enables researchers to detect emerging technological trends, forecast potential breakthroughs, and identify the interdisciplinary nature of innovation.

The integration of various patentometric approaches allows for a comprehensive analysis of innovation ecosystems. Love and Roper (2015), for example, combined patent citations, co-patenting networks, and patent text analysis to assess the impact

of government R&D subsidies on firm-level innovation. This multi-dimensional approach provides a richer understanding of how different factors influence R&D performance and supports the identification of key actors who contribute to technological advancement.

Despite the advantages of patentometrics, certain limitations should be acknowledged. Patent data may not fully capture informal innovation activities, and reliance on patent counts alone can overlook actors who contribute through open innovation or collaborative research without seeking formal intellectual property rights. Therefore, combining patentometrics with alternative indicators such as publication data, funding records, and industry collaborations may provide a more holistic view of innovation dynamics.

Patent Data as a Tool for Studying Technology Development

Patent data serves as a critical resource for studying technology development, offering valuable insights into the processes of invention, diffusion, and commercialization across industries. The systematic analysis of patent data allows researchers to map technological trajectories and identify emerging innovation trends (Cohen et al., 2024). Furthermore, patent data provides insights into collaborative networks and knowledge flows between actors, presenting a holistic view of the innovation ecosystem (Wang et al., 2025).

The utility of patent data spans across various disciplines. Griliches (1990) highlighted its value as an economic indicator, offering insights into firms' productivity and technological capabilities. Similarly, Trajtenberg, Henderson, and Jaffe (1997) employed patent citation analysis to trace the diffusion of knowledge across different sectors, demonstrating the interconnected nature of technological advancements.

In addition to technological insights, patent data can shed light on the geographical distribution of innovation activities. For instance, Carlino, Chatterjee, and Hunt (2007) examined the spatial concentration of patenting activities in the United States, identifying key innovation hubs and the factors contributing to their success. Such geographic analyses assist policymakers and researchers in understanding regional variations in technological development and designing targeted strategies to promote innovation. These insights are particularly relevant in crafting regional innovation policies, ensuring balanced economic growth, and preventing regional disparities in technological development.

Moreover, patent data has proven invaluable in assessing the role of universities and research institutions in technological progress. Studies by Mowery, Nelson, Sampat, and Ziedonis (2004) and Thursby and Thursby (2002) demonstrated the significant role of university patents in fostering industry-academia collaborations and driving technological innovation. These collaborations often facilitate the commercialization of cutting-edge technologies and the emergence of new industries.

Patent data also provides an opportunity to analyze technology life cycles, helping businesses and policymakers identify periods of rapid innovation and subsequent maturity phases. Understanding these patterns allows stakeholders to anticipate

market shifts, allocate resources efficiently, and prioritize research efforts in areas with the highest potential impact.

Furthermore, patent analytics can be used to evaluate cross-sectoral innovation, examining how technologies from different industries converge to create new applications and business opportunities. This interdisciplinary approach is crucial in understanding emerging fields such as artificial intelligence, biotechnology, and clean energy, where technological convergence plays a pivotal role in shaping future developments.

Summary and Research Implications

In summary, the literature on technology development emphasizes the dominance of a select group of key actors who drive innovation and influence industry trajectories. Studies leveraging patentometric methodologies have effectively identified these influential entities, utilizing patent data to provide comprehensive insights into their R&D activities and impact. Various approaches, including patent citations, co-patenting networks, and patent text analysis, have been employed to map technological advancement and reveal collaborative networks that underpin innovation ecosystems.

Furthermore, patent data has been recognized as a valuable tool for tracking technology development, offering a wealth of information on invention processes, market diffusion, and commercialization efforts. The systematic analysis of patent data not only aids in tracing technological trajectories but also enhances the understanding of regional innovation dynamics and the contributions of universities and research institutions.

This review underscores the critical need to incorporate the persistence of R&D output into future analyses to ensure a more holistic evaluation of key actors in technological development. Recognizing actors who consistently contribute to innovation over extended periods is essential for accurately capturing their long-term influence and impact on technological progress.

A Novel Method for Identifying Vibrant Actors in Technology Development

This study introduces a novel approach to identifying vibrant actors in technology development by assessing their performance across two key dimensions: activity and continuity. Given the dynamic nature of R&D performance, which cannot be accurately captured by a single indicator at a fixed point in time, this study proposes an approach that calculates annual indicator values to provide a more comprehensive assessment of performance trends.

To mitigate potential misinterpretations arising from short-term fluctuations, the study employs a sliding window approach, which utilizes a three-year performance span to generate annual indicator values. This approach ensures a more stable and reliable evaluation of actors' sustained contributions over time. To achieve this objective, two key indicators are introduced: the Activity Index and the Continuity Index, which are defined and explained as follows:

Activity index

The activity indicator measures the activity level of actor *i* in a specific field *j* in year *y*. This indicator is measured using a sliding window approach, with a window size of 3 years, counting the research outputs in the filed during 3 years (*y*, *y*-1, and *y*-2). The formula for calculating the activity indicator is as follows:

$$A_i^j(y) = P_i^j(y) + P_i^j(y-1) + P_i^j(y-2) \quad (1)$$

where

y represents the observed year.
i represents the observed actor.
j represents the research field.
$P_i^j(y)$ represents the number of research outputs by actor *i* in field *j* in year *y*.
$A_i^j(y)$ represents the activity indicator for actor *i* in field *j* in year *y*.

Continuity index

The research continuity of actors is a critical indicator of innovation sustainability. In this study, an actor's annual research output is represented by binary values: 0 for no output and 1 for output produced. The cumulative continuous output is then calculated yearly. An actor demonstrating consistent output for three consecutive years is considered to have a high level of continuity. The corresponding calculation formula is explained as follows:

Boolean Variable $B_i^j(y)$

The Boolean variable $B_i^j(y)$ represents whether actor *i* in field *j* has research output in year *y*. The definition of is as follows:

$$B_i^j(y)\begin{cases}1 \ , P_i^j(y) \neq 0\\ 0 \ , P_i^j(y) = 0\end{cases} \quad (2)$$

Where $P_i^j(y)$ represents the research output by actor *i* in field *j* in year *y*. A value of 1 indicates the presence of research output, while 0 indicates its absence.

A. Continuous Research Output Count $n_i^j(y)$

The variable $n_i^j(y)$ captures the number of consecutive years in which actor *i* has research outputs in field *j* up to year y. The initial condition is defined as:

$$n_i^j(y_0) = B_i^j(y_0) \quad (3)$$

This implies that the consecutive research output for the initial year y_0 is equivalent to the value of $B_i^j(y_0)$. For subsequent years, $n_i^j(y)$ is determined as follows:

$$n_i^j(y) = \begin{cases} n_i^j(y-1) + B_i^j(y) \ , if B_i^j(y) \neq 0 \\ 0 \ , if B_i^j(y) = 0 \end{cases} \quad (4)$$

Where

y_0 Initial year of observation.

$B_i^j(y)$ Boolean variable indicating whether research output was in year y.

Continuity Indicator $C_i^j(y)$

To capture broader trends in research output continuity, we define the continuity indicator $C_i^j(y)$, which incorporates a sliding window of three years (SW=3). The formula for calculating the continuity indicator is as follows:

$$C_i^j(y) = n_i^j(y) + \frac{n_i^j(y-1) + n_i^j(y-2)}{2} \quad (5)$$

Where:

$C_i^j(y)$ represents the continuity indicator for actor *i* in field *j* in year *y*.

$n_i^j(y)$ represents the number of consecutive years of research outputs.

This formulation balances recent activity $n_i^j(y)$ with the historical continuity of the preceding two years $n_i^j(y-1)$ and $n_i^j(y-2)$.

Identifying Vibrant Actors

This study introduces the concept of vibrant actors, who must exhibit activity and continuity in R&D output that surpass the average performance of all actors within the field. Therefore, they must meet the following three conditions:

1. The activity index of actor *i* in field *j* during year *y* must be greater than the average activity index of all actors in year *y*. Formula is as follows:

   $$A_i^j(y) > \frac{\sum_{i=1}^{I} A_i^j(y)}{I(y)} = \overline{A_i^j(y)} \quad (6)$$

   Where *I*(*y*) represents the total number of actors in year *y*.

2. The continuity index of actor *i* in field *j* during year *y* must be greater than the average continuity index of all actors in year *y*. Formula is as follows:

$$C_i^j(y) > \frac{\sum_{i=1}^{I} C_i^j(y)}{I(y)} = \overline{C_\iota^J(y)} \quad (7)$$

3. The conditions (1) and (2) must be satisfied for at least three consecutive years (SW=3). Formula is as follows:

$$\forall y \in [y_0, y_0 + sw - 1], \left(A_i^j(y) > \overline{A_\iota^J(y)}\right) \cap \left(C_i^j(y) > \overline{C_\iota^J(y)}\right) \quad (8)$$

This indicates that during the period from y_0 to $y_0 + SW$-1, both $A_i^j(y)$ and $C_i^j(y)$ must be greater than their respective averages, and this condition must be met for at least $y + SW - 1$ consecutive years.

The following example illustrates the process of selecting vibrant assignees in this study, as shown in Table 1. The annual number of patent applications filed by Assignee i is represented as $P_i^j(y)$, from which the annual Activity performance values $A_i^j(y)$ can be calculated. Comparing these values with the average Activity performance of all assignees in the field, $\overline{A_\iota^J(y)}$, it can be observed that Assignee i's $A_i^j(y)$ values exceed the average $\overline{A_\iota^J(y)}$ in the year.

Regarding Continuity performance, the Boolean value $B_i^j(y)$ indicates whether Assignee i produced patents in a given year, while the cumulative number of consecutive years with patent applications is represented as $n_i^j(y)$. Using a three-year performance span, the annual Continuity performance values $C_i^j(y)$ can be calculated. Comparing these values with the average Continuity performance of all assignees in the field, $\overline{C_\iota^J(y)}$, it can be observed that Assignee i's $C_i^j(y)$ values exceed the average $\overline{C_\iota^J(y)}$ in the year.

Finally, when evaluating whether Assignee i consistently meets the threshold of having both $A_i^j(y)$ and $C_i^j(y)$ values above the field average for at least three consecutive years, it is found that Assignee i satisfies this requirement during the periods 2003–2005 and 2014–2020. Therefore, this study identifies Assignee i as a vibrant assignee based on its performance.

Table 1. Sample of a Vibrant Assignee.

	'02	*'03*	*'04*	*'05*	*'06*	*'07*	*'08*	*'09*	*'10*	*'11*	*'12*	*'13*	*'14*	*'15*	*'16*	*'17*	*'18*	*'19*	*'20*	*'21*
$P_i^j(y)$	5	2	1	0	0	0	0	1	1	0	0	1	1	5	1	3	2	1	0	0
$A_i^j(y)$	**5**	**7**	**8**	**3**	1	0	0	1	**2**	**2**	1	1	**2**	**7**	**7**	**9**	**6**	**6**	**3**	1
$B_i^j(y)$	1	1	1	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	0	0
$n_i^j(y)$	1	2	3	0	0	0	0	1	2	0	0	1	2	3	4	5	6	7	0	0
$C_i^j(y)$	1	**2.5**	**4.5**	**2.5**	**1.5**	0	0	**1**	**2.5**	**1.5**	**1**	**1**	**2.5**	**4.5**	**6.5**	**8.5**	**10.5**	**12.5**	**6.5**	**3.5**
$\overline{A_\iota^J(y)}$	1.57	1.73	1.63	1.46	1.35	1.41	1.24	1.38	1.58	1.36	1.09	1.13	1.32	1.54	1.63	1.65	1.81	2.13	1.84	1.39
$\overline{C_\iota^J(y)}$	1.00	1.05	0.93	1.02	1.13	1.12	0.94	0.84	1.21	1.29	0.89	0.90	0.93	1.08	1.09	1.15	1.23	1.46	1.09	0.89

An Empirical Study on Solid-State Electrolyte Technology for Lithium Batteries

This study focuses on solid-state electrolyte technology for lithium batteries as the subject of its empirical analysis, recognizing its transformative impact on battery innovation. Solid-state electrolyte technology addresses critical limitations of conventional lithium-ion batteries, particularly in terms of environmental sustainability and safety, positioning it as a key driver of technological progress and sustainable development. With the global push for carbon neutrality and the growing demand for renewable energy—especially in applications such as electric vehicles and energy storage systems—solid-state electrolyte technology is emerging as a crucial enabler. It enhances battery performance and safety, minimizes environmental impact, and accelerates the adoption of green technologies, thereby supporting global emission reduction targets and advancing renewable energy initiatives (Li et al., 2022).

By analyzing the R&D activities of vibrant actors within this domain, this study aims to identify the key contributors driving technological advancements. Furthermore, it categorizes the various subfields within solid-state electrolyte technology and conducts an in-depth patent analysis to examine the technological strategies adopted by leading companies. This comprehensive approach provides valuable insights into future development trajectories and the evolving competitive landscape of this pivotal technology.

Patent Data Collection

The patent data used in this study was sourced from the United States Patent and Trademark Office (USPTO). Following the research framework of Karabelli, Birke, and Weeber (2021) and the definition of CPC codes (Cooperative Patent Classification, 2024), the study focused on patents related to solid-state electrolyte technology for lithium batteries. Using the query string:

> *@AD>=20020101<=20211231 AND ((lithium OR ion*) AND solid* AND electrolyte*) AND (H01M10/052*.CPC. AND H01M10/056*.CPC.) NOT (Y02E60/50.CPC. OR H01M10/0563.CPC. OR H01M10/0566.CPC. OR H01M10/0567.CPC. OR H01M10/0568.CPC. OR H01M10/0569.CPC.)

Patents with application dates from 2002 to 2021 were retrieved, resulting in a total of 2,690 patents.

The patent search and filtering process was conducted systematically to identify relevant patents related to lithium battery solid-state electrolytes. Initially, patent data was retrieved from the United States Patent and Trademark Office (USPTO) database, yielding a total of 2,690 patents. To refine the dataset, a set of filtering criteria was applied to ensure the selection of patents closely aligned with the research focus. The filtering process involved examining the independent claims to determine whether the solid-state electrolyte was explicitly mentioned and verifying its application in lithium batteries. In cases where the independent claims did not provide explicit information, the patent specifications were reviewed to assess whether they described technological advancements related to solid-state

electrolytes. Through this rigorous filtering process, a total of 981 relevant patents were identified, representing 223 assignees. These selected patents provide a robust foundation for the subsequent analyses in this study. These patents were further categorized into five sub-fields within solid-state electrolyte technology for lithium batteries.

Patent Activity and Vibrant Assignee Distribution across Solid-State Electrolyte Subfields

As summarized in Table 2, the analysis of patenting activity across five sub-technical fields of solid-state electrolytes for lithium batteries provides valuable insights into the distribution of patents, assignees, and vibrant assignees within each category. Among these subfields, organic polymer solid electrolytes stand out as the most extensively studied, with a total of 293 patents and 126 assignees. This reflects substantial research and commercialization efforts in this domain. However, despite the high level of activity, the proportion of vibrant assignees—those demonstrating sustained and impactful R&D contributions—remains at only 10%, underscoring the need for persistent innovation to maintain competitiveness in this rapidly evolving field.

Table 2. Summary of Solid-state Electrolytes for Lithium Batteries.

sub-technical field	***Patents***	***Assignees***	***Vibrant Assignees***
Halides solid electrolytes	134	30	4(0.13)
Mixed inorganic/organic solid electrolytes	99	48	4(0.08)
Organic polymers solid electrolytes	293	126	12(0.10)
Oxides solid electrolytes	236	80	8(0.10)
Sulfide solid electrolytes	219	62	7(0.11)

In contrast, halide solid electrolytes, with 134 patents and 30 assignees, exhibit the highest vibrant assignee ratio at 13%. This indicates a more concentrated distribution of key contributors who consistently drive technological progress. Despite having a lower overall patent volume, this field benefits from a dedicated group of persistent innovators, highlighting the strategic importance of long-term R&D commitment in advancing the technology.

Conversely, mixed inorganic/organic solid electrolytes have the lowest proportion of vibrant assignees at 8%, with 99 patents and 48 assignees. This indicates a more fragmented innovation landscape, where numerous entities contribute to the field, but relatively few sustain a long-term, high-impact presence. The lower ratio of vibrant assignees suggests that consistent innovation efforts are less prevalent in this category, potentially hindering the field's long-term development trajectory and competitiveness.

Oxide and sulfide solid electrolytes, with 236 and 219 patents respectively, demonstrate similar characteristics, exhibiting vibrant assignee ratios of 10% and 11%. These fields strike a moderate balance between the volume of patents and the

persistence of key players, indicating steady and ongoing contributions to technological advancement. Notably, the slightly higher vibrant assignee ratio for sulfide solid electrolytes suggests a more committed group of researchers and institutions, which may further drive consistent progress in this domain.

Overall, the data highlights the critical role of sustained R&D engagement in fostering meaningful and lasting contributions to technological development. While certain subfields, such as halide solid electrolytes, exhibit a strong core of persistent innovators, others—particularly mixed inorganic/organic solid electrolytes—show a broader distribution of participants but may benefit from a more concentrated focus on long-term research efforts. These insights offer valuable guidance for stakeholders aiming to identify key areas of opportunity and strategically invest in the future of lithium battery solid-state electrolyte technologies.

Identifying Vibrant Assignees in Organic Polymers Solid Electrolytes

The persistence of R&D activity among vibrant assignees in the organic polymer solid electrolyte sub-technical field from 2002 to 2021 is a critical aspect of this study, as summarized in Table 3. This analysis captures two key indicators—A values and C values, representing different dimensions of innovation performance. Notably, the C value is of particular significance, as it reflects the sustained and cumulative impact of an assignee's R&D efforts over time. The boxed periods in the table highlight instances where both A and C values exceeded 1 for at least three consecutive years, providing clear evidence of persistent, long-term contributions—one of the core focuses of this research.

A key finding from the data is that vibrant assignees consistently achieve higher and more sustained C values over time compared to their non-vibrant counterparts. For instance, Samsung Electronics and LG Energy Solution, two of the most prominent vibrant assignees, display consistently high C values across multiple years, with extended boxed periods indicating a strong, continuous impact on technological development. These companies not only achieve notable innovation output in specific years but also maintain a steady pace of impactful contributions over the long term. Their persistence underscores a strategic commitment to R&D and an ability to continuously innovate, reinforcing their position as key players in the organic polymer solid electrolyte sector.

In contrast, non-vibrant assignees, despite holding a higher number of patents, often demonstrate fluctuating and less sustained C values, suggesting that their contributions are more sporadic and reactive rather than proactive. For example, assignees such as General Motors and Hydro-Quebec, while possessing relatively high patent counts, lack the consistent upward trend in C values observed among vibrant assignees. Their intermittent bursts of activity, without sustained periods of high C values, suggest that their influence on the technological development of solid-state electrolytes may be transient rather than enduring. This distinction highlights the critical role of persistence—while patent quantity is important, the true measure of technological influence lies in consistent, long-term contributions, as evidenced by the sustained C values of vibrant assignees.

Table 3. Performance of Vibrant and Non-vibrant Assignees in Organic polymers solid electrolytes.

Application Year		*'02*	*'03*	*'04*	*'05*	*'06*	*'07*	*'08*	*'09*	*'10*	*'11*	*'12*	*'13*	*'14*	*'15*	*'16*	*'17*	*'18*	*'19*	*'20*	*'21*
BOSCH (12)	A											0.92	0.88	0.76	0.65	0.62	1.21	3.31	4.69	4.89	2.88
	C											1.12	0.56	0.54	0.93	0.46	1.3	2.03	3.08	2.29	1.69
CNRS (5)	A										0.74	0.92	0.88		1.3	1.85	2.42	1.1	0.47		
	C										0.78	0.56	0.56		0.93	2.3	3.91	2.03	1.03		
HITACHI (5)	A		0.58	0.61	2.05	2.22	2.13	0.81		0.63	0.74	0.92									
	C		0.96	0.54	1.47	2.22	1.34	1.06		0.83	0.39	0.56									
HON HAI PRECISION (3)	A											1.84	2.65	2.27	0.65						
	C											1.12	2.78	1.62	0.93						
LG ENERGY SOLUTION (21)	A							1.62	1.45	1.26				3.02	3.9	3.69	2.42	6.06	6.1	5.97	1.44
	C							1.06	0.59	0.41				1.08	2.32	1.38	1.74	2.03	3.08	2.29	1.69
MURATA (3)	A													0.76	1.3	1.85	1.21	0.55			
	C													1.08	2.32	4.14	2.17	1.22			
NIPPON SODA (5)	A		0.58	1.22	2.05	2.22	1.42	0.81			0.74	0.92	0.88								
	C		0.96	2.68	4.42	5.78	3.13	2.13			0.78	0.56	0.56								
NISSHINBO (3)	A	0.64	1.16	1.22	1.37	0.74	0.71														
	C	1	2.39	1.61	1.96	0.44	0.45														
NITTO DENKO (7)	A		0.58	0.61	1.37	1.48	1.42	1.62	2.18	1.89	1.47	0.92	0.88	0.76							
	C		0.96	0.54	1.47	2.22	1.34	2.13	2.96	1.24	0.78	1.12	0.56	0.54							
SAMSUNG ELECTRONICS (24)	A	3.18	4.05	4.9	2.05	0.74			0.73	1.26	1.47	0.92	0.88	1.51	4.55	4.31	5.45	3.31	2.81	1.63	0.72
	C	1	2.39	4.82	2.45	1.33			1.19	2.07	1.17	1.12	1.11	2.7	4.18	5.98	7.38	8.52	8.56	5.95	3.95
SANYO ELECTRIC (4)	A			1.22	1.37	2.22	1.42	1.62	0.73												
	C			1.07	0.49	1.33	2.24	1.59	1.19												
SEEO (14)	A							0.81	2.18	2.53	2.95	1.84	1.76	1.51	1.3	3.08	3.03	3.31	1.41	1.09	0.72
	C							1.06	2.96	3.72	5.06	3.93	3.33	2.7	1.39	1.84	2.17	3.65	4.45	3.2	2.25
HYDRO-QUEBEC (9)	A	1.27	2.32	2.45	1.37			0.81	0.73	0.63			0.88	0.76	1.30	0.62	1.21	1.10	0.94	0.54	
	C	1.00	2.39	1.61	0.98			1.06	0.59	0.41			1.11	0.54	1.39	0.46	1.30	2.03	1.03	0.92	
UNIVERSITY OF CALIFORNIA (7)	A	1.27	1.16	1.22	0.68	0.74	1.42	0.81	0.73			0.92	0.88	1.51	0.65	1.23	0.61	0.55			
	C	1.00	0.48	0.54	0.98	0.44	1.34	0.53	0.59			1.12	0.56	1.62	0.46	1.38	0.43	0.41			
KUREHA (6)	A			0.61	0.68	0.74		0.81	0.73	0.63	0.74	0.92	0.88		0.65	0.62	1.21	0.55	0.94	0.54	0.72
	C			1.07	0.49	0.44		1.06	0.59	0.41	0.78	0.56	0.56		0.93	0.46	1.30	0.41	1.03	0.46	0.56
COMMISSARIAT A L'ENERGIE ATOMIQUE (5)	A													1.51	1.30	1.23	1.82	1.65	1.41		
	C													1.08	0.46	0.46	0.87	0.41	0.34		
GENERAL MOTORS (5)	A											0.92	0.88	0.76		0.62	0.61	0.55	1.41	1.63	2.16
	C											1.12	0.56	0.54		0.92	0.43	0.41	0.68	0.46	0.56

*__A:__ The value of $A_i^j(y)/\overline{A_i^J(y)}$ for the assignee.

**__C:__ The value of $C_i^j(y)/\overline{C_i^J(y)}$ for the assignee.

***Patent Count of the assignee.

****__Italicized assignee name:__ The top 2 non-vibrant assignees by patent count.

*****__Boxed:__ The period during which both A and C values were greater than 1 for at least three consecutive years.

Furthermore, the analysis reveals that even among vibrant assignees, the timing and duration of high C values vary, offering insights into different innovation strategies.

Some companies, such as SEEO, demonstrate a late but steady rise in C values, indicating their evolving role within the field. This trend suggests that certain companies may transition from being non-vibrant to vibrant assignees by gradually increasing their sustained impact over time, reinforcing the importance of monitoring persistence as an indicator of future influence.

Another key observation is that the relationship between patent counts and sustained impact is not always direct. Some vibrant assignees with relatively fewer patents, such as Murata and Hon Hai Precision, exhibit strong C value performance over multiple years, emphasizing their focus on high-impact, enduring innovations rather than high-volume patenting strategies. This finding underscores the significance of persistence over sheer quantity in assessing an assignee's long-term technological footprint.

Overall, the findings reinforce the central argument of this study—persistent innovation efforts, as captured through high and sustained C values, provide a more accurate reflection of an assignee's true influence in the organic polymer solid electrolyte sector. Vibrant assignees distinguish themselves not merely by patent output but by their ability to maintain a consistent and meaningful presence in the technological landscape over time. These insights offer valuable guidance for policymakers, investors, and industry stakeholders in identifying and supporting long-term contributors to innovation, ensuring that resources are directed towards entities that demonstrate sustained, impactful R&D efforts.

Conclusion

This study presents a novel framework for identifying vibrant actors in technological development by emphasizing the persistence of their R&D efforts alongside their overall activity. The findings reveal that traditional patentometric approaches—primarily focused on patent counts and citation frequencies—often fail to capture the critical dimension of sustained innovation. By developing and applying the activity and continuity indices, this study effectively distinguishes vibrant assignees, who demonstrate consistent and impactful contributions over time, from non-vibrant assignees, who may achieve high patent output but lack sustained engagement.

The empirical analysis of solid-state electrolyte technology for lithium batteries further reinforces the importance of persistence in driving technological advancements. The results indicate that vibrant assignees, such as Samsung Electronics and LG Energy Solution, consistently achieve high continuity values, reflecting their long-term commitment to R&D and strategic positioning within the industry. Conversely, non-vibrant assignees, despite holding extensive patent portfolios, often exhibit fluctuating continuity values, suggesting sporadic involvement and a lack of sustained impact.

These findings offer valuable insights for policymakers, industry stakeholders, and investors, helping them better identify and support key contributors to technological innovation. By integrating persistence as a core factor in R&D evaluation, decision-makers can optimize resource allocation, foster strategic partnerships, and strengthen the overall innovation ecosystem.

Moreover, the proposed framework provides a more comprehensive approach to tracking technological leadership and identifying potential emerging players. By considering both the frequency and sustainability of contributions, this approach offers a deeper understanding of innovation dynamics, supporting more informed decision-making processes in policy and investment planning.

Future Research

Future research can build upon the proposed framework by applying it across various industries to examine the persistence of R&D efforts in diverse technological landscapes. While this study focuses on solid-state electrolytes for lithium batteries, the methodology can be effectively adapted to other high-impact sectors. Analyzing the sustained performance of key players in different industries can provide deeper insights into how persistent innovation drives long-term technological leadership and competitiveness.

Beyond corporate assignees, the framework can be extended to academic research institutions by analyzing journal articles and publication data. Evaluating the sustained contributions of universities and research organizations can offer valuable insights into their research impact and long-term influence across scientific domains. This extension can assist funding agencies, policymakers, and institutional leaders in better understanding and fostering innovation within the academic ecosystem, ultimately guiding strategic decision-making and resource allocation.

Additionally, the framework can be refined to assess individual-level vibrant performance, focusing on inventors and authors. By tracking personal research trajectories based on persistence in patenting or publishing, it becomes possible to identify prolific innovators and thought leaders who consistently contribute to technological and scientific advancements. Such insights can support talent management strategies, facilitate targeted collaborations, and help organizations recognize and retain top-performing researchers.

Expanding the application of this framework across industries, academic institutions, and individual contributors will not only enhance its versatility but also provide a more comprehensive understanding of innovation ecosystems. Future research efforts can focus on developing sector-specific benchmarks, refining the methodology to accommodate discipline-specific nuances, and leveraging advanced analytics to further enhance the precision and applicability of vibrant performance evaluations.

Acknowledgments

This work was financially supported by the Center for Research in Econometric Theory and Applications (Grant no. 113L900202) which is under the Featured Areas Research Center Program by Higher Education Sprout Project of Ministry of Education (MOE) in Taiwan.

References

Blind, K. (2012). The Influence of Regulations on Innovation: A Quantitative Assessment for OECD Countries. *Research Policy*, *41*(2), 391-400.

Breschi, S., Malerba, F., & Orsenigo, L. (2000). Technological Regimes and Schumpeterian Patterns of Innovation. *The Economic Journal, 110*(463), 388-410.

Carlino, G. A., Chatterjee, S., & Hunt, R. M. (2007). Urban Density and the Rate of Invention. *Journal of Urban Economics, 61*(3), 389-419.

Cohen, M., Fernandes, G., & Godinho, P. (2024). Measuring the Impacts of University-Industry R&D Collaborations: A Systematic Literature Review. *The Journal of Technology Transfer*, 1-30.

Cooperative Patent Classification. (2024). CPC scheme and CPC definitions. Retrieval: https://www.cooperativepatentclassification.org/cpcSchemeAndDefinitions/table

Dolphin, G., & Pollitt, M. (2020). Identifying Innovative Actors in the Electricicity Supply Industry Using Machine Learning: An Application to UK Patent Data (No. 2013). Faculty of Economics, University of Cambridge.

Fleming, L., & Sorenson, O. (2004). Science as a Map in Technological Search. *Strategic Management Journal, 25*(8-9), 909-928.

Grilliches, Z. (1990). Patent Statistics as Economic Indicators: A Survey Part I. *NBER Working Paper, 3301*(Part I).

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2002). Market Value and Patent Citations: A First Look (No. 0201001). University Library of Munich, Germany.

Huang, Z., Notten, A., & Rasters, N. (2011). Nanoscience and Technology Publications and Patents: A Review of Social Science Studies and Search Strategies. *The Journal of Technology Transfer*, 36(2), 145-172.

Karabelli, D., Birke, K. P., & Weeber, M. (2021). A Performance and Cost Overview of Selected Solid-State Electrolytes: Race between Polymer Electrolytes and Inorganic Sulfide Electrolytes. *Batteries, 7*(1), 18.

Lerner, J., & Seru, A. (2017). The Use and Misuse of Patent Data: Issues for Corporate Finance and Beyond. *NBER Working Paper Series*, No. 24022.

Li, M., Lu, J., Chen, Z., & Amine, K. (2022). 30 Years of Lithium-Ion Batteries. *Advanced Materials, 30*(33), 1800561.

Love, J. H., & Roper, S. (2015). SME Innovation, Exporting and Growth: A Review of Existing Evidence. *International Small Business Journal*, 33(1), 28-48.

Mowery, D. C., Nelson, R. R., Sampat, B. N., & Ziedonis, A. A. (2015). *Ivory Tower and Industrial Innovation: University-Industry Technology Transfer Before and After the Bayh-Dole Act*. Stanford University Press.

Narin, F., Noma, E., & Perry, R. (1987). Patents as Indicators of Corporate Technological Strength. *Research Policy, 16*(2-4), 143-155.

Thursby, J. G., & Thursby, M. C. (2002). Who is selling the Ivory Tower? Sources of Growth in University Licensing. *Management Science, 48*(1), 90-104.

Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus Corporate Patents: A Window on the Basicness of Invention. *Economics of Innovation and New Technology, 5*(1), 19-50.

Valkokari, K., Amitrano, C. C., Bifulco, F., & Valjakka, T. (2016). Managing Actors, Resources, and Activities in Innovation Ecosystems–A Design Science Approach.

In Collaboration in A Hyperconnected World: 17th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2016, Porto, Portugal, October 3-5, 2016, Proceedings 17 (pp. 521-530). Springer International Publishing.

Wang, Y., Li, Y., Ding, P., & Guo, B. (2025). Technology Transfer and Innovation Efficiency in a Large Emerging Economy: An Integrative Perspective of Absorptive Capacity and the Technology Ladder. *The Journal of Technology Transfer*.

https://doi.org/10.51408/issi2025_053

Impact of Web of Science and Scopus Policies on Multiple Document-Type Classification

Domenico A. Maisano[1], Lucrezia Ferrara[2], Fiorenzo Franceschini[3]

[1]*domenico.maisano@polito.it*, [2]*lucrezia.ferrara@polito.it*, [3]*fiorenzo.franceschini@polito.it*
Politecnico di Torino, DIGEP (Department of Management and Production Engineering),
Corso Duca degli Abruzzi 24, 10129, Torino (Italy)

Abstract

Document-type (DT) classification – i.e., the assignment of conventional labels such as *article*, *review*, *proceedings paper*, etc., to scientific documents – is crucial for information retrieval in bibliometric databases, but its incomplete objectivity can lead to errors with implications on indicators and research evaluations. This study focuses on a portion of the documents (with a relatively small incidence ~4%) with dual-DT assignment in Web of Science (WoS) – a feature that is absent in Scopus, which applies only single-DT assignments – to assess their characteristics and classification accuracy.

A manual analysis of more than a thousand documents revealed three main scenarios of dual-DT assignment in WoS: (i) the combination of one DT describing the *content* and another describing the *container* (e.g., *book chapters*, *proceedings papers*), (ii) the handling of specialized DTs (e.g., *data paper*, *retracted* paper), and (iii) the combination of a DT related to journal publication with a temporary DT for the *early-access* designation.

Documents with dual-DT assignment in WoS exhibit higher error rates, confirming the greater difficulty of classification for both databases, even for Scopus, regardless of its single-DT policy. WoS's dual-DT classification policy offers more detail and potentially greater accuracy but also shows some inconsistencies. Conversely, Scopus's single-DT policy reduces the level of detail and increases the risk of misclassification, particularly for papers from *conference proceedings* or journal *special issues*.

This study highlights the need for clearer DT definitions and recommends that bibliometric databases consider adopting more flexible multiple-DT classification policies to enhance both detail and accuracy in document classification. A limitation of this research is the relatively small *corpus* of documents analysed, which will be expanded in future studies.

Introduction

Document types (DTs) – such as research *articles*, *reviews*, *proceedings papers*, and *book chapters* – are conventional labels applied to scientific documents to describe their nature and main characteristics, facilitating information retrieval (Donner, 2017; Yeung, 2021). Depending on the publication context, DTs can be assigned by various stakeholders, including authors, editorial boards, publishers and bibliometric databases. However, because there are no universally accepted definitions or standardized rules for DT classification, a degree of subjectivity is unavoidable. This subjectivity often leads to questionable or even erroneous classifications. For instance, a *review* or a *note* might be misclassified as a research *article*, leading to several potential consequences. Beyond misleading researchers during document searches, these classification errors can distort bibliometric indicators for journals, individual researchers, and entire research institutions. Such distortions arise because bibliometric indicators often depend on the DT classification of the documents under

analysis. For instance, if a journal mislabels a substantial number of documents as *articles* rather than, say, *editorials* or *letters*, its *Journal Impact Factor* could be distorted: citations to those misclassified items would still be counted in the numerator, while the denominator (which includes only *articles* and *reviews*) might be inappropriately inflated or deflated (Haupka et al., 2024).

In some cases, these errors may even impact research evaluation exercises, which frequently include or exclude documents based on their DTs. For example, certain DTs – such as *proceedings papers*, *notes*, and *book chapters* – are often deemed less significant and are excluded from evaluations (García-Pérez, 2010; Franceschini et al., 2015; Yeung, 2019; Mokhnacheva, 2023). A *conference paper* erroneously classified as a journal *article* might grant a researcher undue credit in evaluations that prioritize journal publications, potentially influencing hiring, promotion, or funding decisions. Conversely, an important research contribution misclassified as a less prestigious DT (e.g., an *article* mislabelled as a *note*) could be undervalued in performance assessments.

Additionally, different research disciplines may be affected by misclassification of DTs in distinct ways. Fields that make heavy use of conference proceedings (e.g., *computer science* and *engineering*) may be particularly susceptible to misclassification between *conference papers* and journal *articles*, whereas disciplines that focus primarily on journal articles (e.g., *biology* and *medicine*) may be more concerned with distinguishing research *articles* from *reviews* or *editorial materials*.

Scientific literature on DT-classification errors is relatively sparse, primarily because such investigations typically involve samples of only a few hundred or thousand documents, requiring labour-intensive manual analysis. Recent studies suggest that DT-classification errors in general-purpose bibliometric databases, such as Web of Science (WoS) and Scopus, are non-negligible and account for a few percentage points (Franceschini et al., 2016a; Yeung, 2021; Donner, 2023; Zhu et al., 2024). These findings are corroborated by a recent study by Maisano et al. (2025), which introduces a semi-automated approach to detect potentially misclassified documents. This approach utilizes discrepancies between DT classifications assigned by competing databases, WoS and Scopus, to automatically identify subsets of potentially misclassified documents. Manual analysis, which is inherently time-consuming, can then be concentrated on this subset while excluding most documents, which are presumed to be correctly classified. This approach allows for an approximately two-order-of-magnitude increase in the size of analysed samples – e.g., from a few thousand to hundreds of thousands – without requiring additional manual-analysis effort. Maisano et al. (2025) analysed a sample of nearly 28,000 documents recently published by over 2,000 researchers affiliated with the two largest universities in Turin, Politecnico di Torino (PoliTO) and Università di Torino (UniTO). The study estimated overall error rates of approximately 2.3% for WoS and 2.7% for Scopus.

During the data collection for the research in (Maisano et al., 2025), an intriguing fact emerged: while most documents indexed by WoS and Scopus featured a single-DT classification, approximately 4% exhibited dual-DT classifications in WoS –

e.g., documents classified simultaneously as *editorial material; book chapter* or *article; proceedings paper* – whereas Scopus consistently applied single-DT classifications. These dual-DT classifications in WoS were excluded from the earlier study to avoid complicating the analysis.

It is important to clarify that in a dual-DT classification by WoS, up to two DTs in combination can be assigned to a single document. The most frequent combination involves one DT indicating the *content* type of the document (e.g., *article, review, letter*) and the other indicating the corresponding publication *container* (e.g., *proceedings paper, book chapter, journal*). However, other combinations of DTs are also possible. In contrast, Scopus' policy limits each document to a single DT label, forcing a choice even in cases where multiple DTs would be appropriate. This conceptual distinction is important, as some DTs are not mutually exclusive; in other words, sometimes a single document may legitimately fall under two DTs.

Building upon these observations, this study specifically focuses on this portion of documents with dual-DT classifications in WoS. The objectives are twofold: (i) to explore the reasons behind WoS's dual-DT assignments, likely indicative of greater classification challenges for these documents, and (ii) to compare WoS's dual-DT-assignment policy with Scopus's single-DT-assignment approach. Formally, the study addresses the following research questions, respectively:

RQ#1: *Is the error rate (for both WoS and Scopus) higher for documents with dual-DT assignments compared to those with single-DT assignments, confirming that the former are inherently more challenging to classify?*

RQ#2: *Based on the analysis, which approach – WoS's dual-DT assignments or Scopus's single-DT assignments – appears more reasonable?*

Methodologically, the study will conduct an exhaustive manual analysis of a *corpus* of documents of interest, assessing the accuracy of DT classifications in WoS and Scopus and attributing errors where detected. The remainder of this study is organized in three sections. The "Methodology" section details the methodological approach, including the sample selection, manual analysis procedure, and statistical measures to be constructed. The "Results" section presents the findings and relevant statistics, accompanied by descriptions, interpretations of the results, and practical examples. Finally, the "Conclusions" section summarizes the key findings, highlights practical implications for the scientific community, discusses limitations, and suggests directions for future research.

Methodology

As outlined in the "Introduction", this study builds on the dataset used in Maisano et al. (2025), which combines publications authored by researchers affiliated with UniTO (a generalist university in Turin) and PoliTO (a technical university in Turin) during the 2019–2023 period. The choice of these two medium-to-large universities – with a combined total of over 100,000 students and approximately 2,000 tenured researchers, covering a wide range of scientific disciplines – ensures that the dataset of publications is diverse in terms of subjects, DTs, journals and publishers, making

it relatively representative of the entire recent scientific literature. These publications are indexed by both WoS and Scopus.

From an initial set of nearly 30,000 documents, 1,085 were identified as having dual-DT assignments by WoS. These documents constitute the *corpus* under investigation in this study. Table 1 provides a detailed classification of these documents by both databases. Notably, the DT labels assigned by WoS and Scopus do not always align, due to minor differences in DT naming conventions (e.g., *conference paper* in Scopus versus *proceedings paper* in WoS) and the inclusion or exclusion of certain specialized DTs (e.g., *expression of concern* and *meeting abstract* in WoS but not in Scopus). For further details, refer to the official DT lists provided by Scopus and WoS (Clarivate, 2025; Elsevier, 2025).

Table 1. Summary of DTs classified by WoS and Scopus for the 1,085 publications analysed in this study. DTs in each database are sorted in descending order based on the number of documents they include.

(a) DTs classified by WoS	No. of docs	(b) DTs classified by Scopus	No. of docs
Article; Proceedings paper	423	*Article*	767
Article; Early access	394	*Book chapter*	152
Article; Book chapter	146	*Conference paper*	54
Article; Data paper	45	*Review*	53
Review; Early access	44	*Data paper*	40
Editorial material; Book chapter	10	*Letter*	7
Letter; Early access	7	*Editorial*	6
Editorial material; Early access	5	*Erratum*	4
Review; Book chapter	5	*Note*	1
Correction; Early access	4	*Retracted*	1
Article; Expression of concern	1		
Article; Retracted publication	1		
Total no. of documents	**1,085**	**Total no. of documents**	**1,085**

Table 2 presents a matrix that highlights the similarities and discrepancies between the DT classifications in WoS (with DTs listed in the rows) and Scopus (with DTs listed in the columns) for the analysed documents. While some classifications appear consistent (e.g., the four documents classified as *correction; early access* in WoS and *erratum* in Scopus), others exhibit clear incompatibilities (e.g., the eight documents classified as *review; early access* in WoS but as *article* in Scopus).

Table 2. Matrix of DT classifications for the analysed documents, according to WoS (rows) and Scopus (columns). All 1,085 documents were manually analysed to identify potential classification errors.

DT classifications →		by Scopus										
	↓	*Article*	*Book chapter*	*Conf. paper*	*Review*	*Data paper*	*Letter*	*Editorial*	*Erratum*	*Note*	*Retracted*	**Row total**
by WoS	*Article; Proceedings paper*	359	-	52	12	-	-	-	-	-	-	**423**
	Article; Early access	389	-	2	3	-	-	-	-	-	-	**394**
	Article; Book chapter	3	143	-	-	-	-	-	-	-	-	**146**
	Article; Data paper	5	-	-	-	40	-	-	-	-	-	**45**
	Review; Early access	8	-	-	36	-	-	-	-	-	-	**44**
	Editorial material; Book chapter	-	7	-	-	-	-	3	-	-	-	**10**
	Letter; Early access	-	-	-	-	-	7	-	-	-	-	**7**
	Editorial material; Early access	1	-	-	-	-		3	-	1	-	**5**
	Review; Book chapter	1	2	-	2	-	-	-	-	-	-	**5**
	Correction; Early access	-	-	-	-	-	-	-	4	-	-	**4**
	Article; Expression of concern	1	-	-	-	-	-	-	-	-	-	**1**
	Article; Retracted publication	-	-	-	-	-	-	-	-	-	1	**1**
	Column total	**767**	**152**	**54**	**53**	**40**	**7**	**6**	**4**	**1**	**1**	**1,085**

All 1,085 documents in the matrix were manually analysed to determine their "true" (or correct) DTs and identify any potential classification errors by the databases. The manual analysis was conducted shortly after data retrieval, in February 2024. Depending on the need, the following information was considered to determine the "true" DT(s) for each document, with a progressively deeper manual analysis where required:

- Title and abstract;
- Information and metadata provided on the journal and/or publisher's webpage;
- Formal structure of the document;
- Number of references cited within the document;
- Full text.

For each document, the accuracy of the DT classification provided by each database was assessed, also considering their respective DT definitions and presumed assignment rules. A logic of internal consistency was applied to establish whether a database's DT classification was correct or erroneous.

Results

This section is divided into two subsections: (i) the presentation of results from the perspective of WoS and Scopus, using so-called "error tables" (Maisano et al., 2025) and associated error statistics, and (ii) the practical interpretation of these results, supported by several pedagogical examples.

Error tables and error statistics

Error! Reference source not found. presents the error table for WoS, a contingency table that displays the DT classifications assigned by WoS in its columns and the "true" (or correct) DTs determined through manual analysis in its rows (Maisano et al., 2025). The main diagonal of the error table contains the correct DT classifications

– with the corresponding document counts shown in round parentheses "(·)" – while the off-diagonal elements represent incorrect classifications. Notably, while the columns include only the DTs listed in Table 1(a) for WoS, additional DTs appear in the last rows of **Error! Reference source not found.**, denoted by the symbol "(*)" – including one with dual-DT classification "*review; proceedings paper*" at the top. These additional DTs were introduced because the manual analysis revealed misclassifications that were corrected by assigning more appropriate DT classifications, following WoS's declared rules (Clarivate, 2025).

Table 3. Error table for WoS. Quantities in "(·)" represent correctly classified documents, "[·]" denote *partial* errors (with weight ½), and "{·}" indicate *full* errors. Statistics α_i and β_j were calculated only for groups with at least 30 documents for statistical reliability. The symbol "(*)" denotes additional DTs added following manual analysis.

"True" DT classifications	DT classification by WoS: *Article; Proceed. paper*	*Article; Early access*	*Article; Book chapter*	*Article; Data paper*	*Review; Early access*	*Editorial material; Book chapter*	*Letter; Early access*	*Editorial material; Early access*	*Review; Book chapter*	*Correction, Early access*	*Article; Expression of concern*	*Article; Retracted publicat.*	Row total	Total "{·}" row errors	Total "[·]" row errors	α_i
Article; Proceedings paper	(409)	[2]	-	-	-	-	-	-	-	-	-	-	411	0	2	**0.2%**
Article; Early access	-	(329)	-	-	-	-	-	-	-	-	-	-	329	0	0	**0.0%**
Article; Book chapter	-	-	(145)	-	-	-	-	-	-	-	-	-	145	0	0	**0.0%**
Article; Data paper	-	-	-	(45)	-	-	-	-	-	-	-	-	45	0	0	**0.0%**
Review; Early access	-	[1]	-	-	(33)	-	-	-	-	-	-	-	34	0	1	**1.5%**
Editorial material; Book chapter	-	-	-	-	-	(8)	-	-	-	-	-	-	8	0	0	-
Letter; Early access	-	-	-	-	-	-	(6)	-	-	-	-	-	6	0	0	-
Editorial material; Early access	-	-	-	-	-	-	-	(3)	-	-	-	-	3	0	0	-
Review; Book chapt.	-	-	-	-	-	-	-	-	(5)	-	-	-	5	0	0	-
Correction, Early access	-	-	-	-	-	-	-	-	-	(4)	-	-	4	0	0	-
Article; Expression of concern	-	-	-	-	-	-	-	-	-	-	(1)	-	1	0	0	-
Article; Retracted publication	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	-
Review; Proceedings paper(*)	[12]	-	-	-	[1]	-	-	-	-	-	-	-	13	0	13	-
Article(*)	-	[58]	-	-	{2}	-	-	-	-	-	-	[1]	61	2	59	-
Review(*)	-	{3}	-	-	[8]	-	-	-	-	-	-	-	11	3	8	-
Letter(*)	-	-	-	-	-	-	[1]	-	-	-	-	-	1	0	1	-
Editorial material(*)	{1}	-	{1}	-	-	-	-	-	-	-	-	-	2	2	0	-
Book chapter(*)	-	-	-	-	-	[2]	-	-	-	-	-	-	2	0	2	-
Proceedings paper(*)	[1]	-	-	-	-	-	-	-	-	-	-	-	1	0	1	-
Other(*)	-	{1}	-	-	-	-	-	{2}	-	-	-	-	3	3	0	-
Column total	423	394	146	45	44	10	7	5	5	4	1	1	1,085			
Total "{·}" column errors	1	4	1	0	2	0	0	2	0	0	0	0		10		
Total "[·]" column errors	13	61	0	0	9	2	1	0	0	0	0	1			87	
β_j	**1.8%**	**8.8%**	**0.7%**	**0.0%**	**14.8%**	-	-	-	-	-	-	-				$\varepsilon \cong$ **4.9%**

Among the off-diagonal elements, two types of errors can be distinguished:

- *Full* errors (quantities denoted in curly brackets "{·}"), representing cases where both DTs assigned by WoS are incorrect.
- *Partial* errors (quantities denoted in square brackets "[·]"), involving cases where one of the two assigned DTs is correct, while the other is incorrect.

Partial errors are weighted with a score of ½, as they represent an intermediate level of error between fully incorrect DT assignments (score of 1) and correct DT assignments (score of 0). The content of the error table can be summarized using an overall (weighted) error rate:

$$\varepsilon = \frac{d^{\{\cdot\}} + ½ \cdot d^{[\cdot]}}{d^{\{\cdot\}} + d^{[\cdot]} + d^{(\cdot)}}, \quad (1)$$

where:

$d^{\{\cdot\}} = \sum_{i,j} d_{i,j}^{\{\cdot\}}$ is the total number of documents with *full* errors in the error table;

$d^{[\cdot]} = \sum_{i,j} d_{i,j}^{[\cdot]}$ is the total number of documents with *partial* errors in the error table;

$d^{(\cdot)} = \sum_{i,j} d_{i,j}^{(\cdot)}$ is the total number of correctly classified documents in the error table.

The denominator of the fraction in Eq. 1 represents the sum of all three document categories, which amounts to 1,085. For WoS, the number of *partial* errors ($d^{[\cdot]} = 87$) is significantly higher than the number of *full* errors ($d^{\{\cdot\}}$=10). Moreover, the error rate for WoS (~4.9%) is markedly higher than the rate observed in the previous study (~2.3%) for documents with single-DT assignments (Maisano et al., 2025). A statistical test on the difference between the two proportions confirmed this rigorously (Ross, 2017). Addressing RQ#1, it can be concluded that documents with dual-DT classifications in WoS are inherently more challenging to classify, as they exhibit a significantly higher propensity for misclassification.

Beyond the overall error rate (ε), additional error statistics can be constructed. Specifically, for each row i, the probability that a document belonging to a given DT is wrongly classified into another DT (i.e., *missing assignment to the DT of interest*) is:

$$\alpha_i = \frac{\sum_j d_{i,j}^{\{\cdot\}} + ½ \cdot \sum_j d_{i,j}^{[\cdot]}}{\sum_j d_{i,j}^{\{\cdot\}} + \sum_j d_{i,j}^{[\cdot]} + \sum_j d_{i,j}^{(\cdot)}}. \quad (2)$$

For each column j, the probability of misclassifying a document into the specific DT of that column (i.e., *false classification into the DT of interest*) is:

$$\beta_j = \frac{\sum_i d_{i,j}^{\{\cdot\}} + ½ \cdot \sum_i d_{i,j}^{[\cdot]}}{\sum_i d_{i,j}^{\{\cdot\}} + \sum_i d_{i,j}^{[\cdot]} + \sum_i d_{i,j}^{(\cdot)}}. \quad (3)$$

These statistics (α_i and β_j) were calculated only for groups with at least 30 total documents (in rows or columns) to ensure statistically reliable estimates. The statistics indicate that several errors involve journal documents with dual-DT assignments including the *early-access* designation, despite being already published in their final form (i.e., with specific volume/issue numbers and definitive page numbers). However, these errors are not particularly severe for two reasons:

1. They do not fundamentally alter the "true" nature of the document; they simply attach an erroneous (temporary) designation of *early access*;

2. From a follow-up investigation conducted approximately 10 months after data collection (December 2024), it was found that over 80% of these inaccuracies had been corrected by WoS.

A smaller proportion of errors involves *conference proceedings papers* misclassified as *articles*, although they are actually *reviews* or *surveys*.
For Scopus (see the relevant error table in
Table **4**), no dual-DT assignments appear, as this database only permits single-DT classifications (Elsevier, 2025). Consequently, only *full* errors are observed, with an error rate of $\varepsilon = \frac{86}{1{,}085} \cong 7.9\%$, significantly higher than the 2.7% reported in the previous study (Maisano et al., 2025). Statistical testing confirmed the significance of this difference. Scopus appears to encounter even greater challenges than WoS when classifying these particularly delicate documents. The distribution of errors in Scopus reveals that many articles are misclassified as *conference papers*, while others classified as *articles* belong to more specific DTs (e.g., *reviews*, *book chapters*, *data papers*). As explored in the next subsection, the root cause of these errors often lies in the limitations of Scopus's DT definitions and its strict single-DT-assignment policy.

Interpretation of results

This subsection provides an interpretation of the most significant results of the analysis, supported by numerous practical examples. Three typical scenarios were observed in which WoS assigns dual-DT classifications, each of which is analysed individually below:

1. Combination of a DT related to a document's *content* and a DT related to the *container* (or dissemination context);
2. *Early-access* documents, typically linked to scientific journals;
3. Classification of uncommon, specialized documents from scientific journal.

Table 4. Error table for Scopus. Quantities in "(·)" represent correctly classified documents, while "{·}" denote *full* errors. Statistics α_i and β_j were calculated only for groups with at least 30 documents for statistical reliability. The symbol "(*)" denotes additional DTs added following manual analysis.

		DT classification by Scopus										Row total	Total row errors	α_i
		Article	*Book chapter*	*Conf. paper*	*Review*	*Data paper*	*Letter*	*Editorial*	*Erratum*	*Note*	*Retracted*			
"True" DT classifications	*Article*	(745)	-	{52}	{4}	-	-	-	-	-	{1}	802	57	**7.1%**
	Book chapter	{3}	(147)	-	-	-	-	-	-	-	-	150	3	**2.0%**
	Conf. paper	-	-	(1)	-	-	-	-	-	-	-	1	0	-
	Review	{11}	{1}	{1}	(48)	-	-	-	-	-	-	61	13	**21.3%**
	Data paper	{5}	-	-	-	(40)	-	-	-	-	-	45	5	**11.1%**
	Letter	-	-	-	-	-	(7)	-	-	-	-	7	0	-
	Editorial	{1}	{4}	-	-	-	-	(6)	-	-	-	11	5	-
	Erratum	-	-	-	-	-	-	-	(4)	-	-	4	0	-
	Note	{2}	-	-	-	-	-	-	-	(1)	-	3	2	-
	Retracted	-	-	-	-	-	-	-	-	-	-	0	0	-
	Short survey(*)	-	-	-	{1}	-	-	-	-	-	-	1	1	-
Column total		767	152	54	53	40	7	6	4	1	1	1,085		
Total column errors		22	5	53	5	0	0	0	0	0	1		86	
β_j		**2.9%**	**3.3%**	**98.1%**	**9.4%**	**0.0%**	-	-	-	-	-			$\varepsilon \cong 7.9\%$

(1) Combination of a DT related to *content* and a DT related to the *container*. A common pairing observed involves (i) a DT describing the document's *content* in terms of objectives and structure (e.g., research *article*, *review*, *letter*), and (ii) a DT related to the *container*, representing the dissemination context (e.g., *journal*, *conference proceedings*, *book chapter*). Specifically, WoS seems to include the container DT only for scientific publications that differ from *journal* contributions[1]. This practice is consistent with the WoS definition of *proceedings paper*, which states, "*proceedings papers will have a dual document type: article; proceedings paper*", though no similar rule exists for *book chapters*, which are defined only as "*a monograph or publication written on a specific topic within a main division in a book*" (Clarivate, 2025). On the other hand, Scopus, constrained by its single-DT-assignment policy, provides systematically less detailed classifications and occasionally misleading ones. The examples below document some of the most common and/or curious errors observed for both databases.

- For example, Scopus defines a *book chapter* as "*a complete chapter in a book or book-series volume, identified as a chapter by a heading or section indicator*". However, some special *book chapters*, such as book series introductions, are often classified by Scopus as *editorial*. Additionally, inconsistencies arise because some book-series *editorials* are still classified as *book chapters*. These internal inconsistencies in Scopus are generally avoided by WoS due to its dual-DT-assignment policy.

  For example, documents 1.1 to 1.5 in Table 5 pertain to *book chapters*. The first three are from the same book but differ in content: the first is an *introduction* to the whole book, the second is a classic research *article* (complete with methodology, results, discussion, etc.), and the third contains concluding notes related to the whole book. WoS assigns the container-DT *book chapter* to all three, pairing it with a content-DT: *article* for the second and *editorial material* for the introduction and conclusions, consistent with its definition of *editorial material* (Clarivate, 2025). Conversely, Scopus classifies the first document as *editorial* (but not as *book chapter*), while the other two are classified as *book chapters* (but not as *article* or *editorial*). Although this DT classification is not exactly wrong, it is undoubtedly less detailed and potentially more misleading than that of WoS.

  Focusing on documents 1.4 and 1.5 in Table 5, both from another book, WoS not only classifies them correctly as *book chapters* but also distinguishes their content by assigning the additional DTs *review* and *article*, respectively. In contrast, Scopus's classification, while accurate, assigns only the single DT *book chapter* to both contributions.

[1] In fact, the DT *journal* is not envisaged by WoS as it is implied when the database makes a single-DT assignment related to the document's content, such as *article*, *review* or *letter*.

Table 5. Examples of documents with peculiar DT classifications, discussed in the analysis.

Ref.	DOI	DT classification		Brief description
		WoS	Scopus	
1.1	https://doi.org/10.1007/978-3-319-71837-8_1	*Editorial material; book chapter*	*Editorial*	Introductory chapter of a book divided into chapters.
1.2	https://doi.org/10.1007/978-3-319-71837-8_10	*Article; book chapter*	*Book chapter*	Chapter corresponding to a research *article*.
1.3	https://doi.org/10.1007/978-3-319-71837-8_12	*Editorial material; book chapter*	*Book chapter*	Concluding chapter of a book divided into chapters.
1.4	https://doi.org/10.1007/978-3-319-79084-8_2	*Review; book chapter*	*Book chapter*	Same as 1.2, but corresponding to a *review*.
1.5	https://doi.org/10.1007/978-3-319-79084-8_3	*Article; book chapter*	*Book chapter*	Same as 1.2.
2.1	https://doi.org/10.1111/odi.13076	*Article; proceedings paper*	*Conf. paper*	*Review* published in a journal special issue dedicated to a medical workshop (7th WWOM).
3.1	https://doi.org/10.5004/dwt.2018.22308	*Article; proceedings paper*	*Article*	*Article* in a journal special issue dedicated to an international conference (CEST 2017).
3.2	https://doi.org/10.5004/dwt.2018.22995	*Article; proceedings paper*	*Article*	Same as 3.1.
3.3	https://doi.org/10.5004/dwt.2019.24424	*Article; proceedings paper*	*Article*	*Article* in a journal special issue dedicated to an international conference (NAXOS 2018).
3.4	https://doi.org/10.5506/APhysPolB.51.1627	*Article; proceedings paper*	*Article*	*Article* in a journal (Acta Physica Polonica B) exclusively dedicated to conference proceedings.
3.5	https://doi.org/10.5506/APhysPolB.51.655	*Article; proceedings paper*	*Article*	Same as 3.4.
3.6	https://doi.org/10.5506/APhysPolB.51.661	*Article; proceedings paper*	*Article*	Same as 3.4.
3.7	https://doi.org/10.1200/JCO.18.00053	*Article; proceedings paper*	*Article*	In the "Prior presentation" section of this journal *article*, it is stated that the contribution was presented in three different conferences held in 2017 and 2018.
4.1	https://doi.org/10.1002/bmb.21490	*Editorial material; early access*	*Article*	Christmas song called "*Oxidosqualene (OS) cyclase—Lanosterol synthase*", appearing in the scientific journal (Biochemistry and Molecular Biology Education), which classifies the paper with the specialized DT *the lighter side*.
4.2	https://doi.org/10.1002/alz.13526	*Article; early access*	*Article*	In the "Presentations" section of this journal *article*, it is stated that the contribution was presented at three different conferences held in 2021 and 2023. Following its final publication, WoS's *early-access* designation was removed, leaving the single DT *article*.
4.3	https://doi.org/10.1002/hon.3184	*Article; early access*	*Article*	Before the abstract, it is stated that "*Preliminary results were presented as an abstract and oral presentation at the 63rd ASH Annual Meeting & Exposition in 2021*". Following its final publication, WoS's *early-access* designation was removed, leaving the single DT *article*.
5.1	https://doi.org/10.1002/ecy.2448	*Article; data paper*	*Article*	Document published in a journal as a *data paper*, containing detailed information on the dataset used for a research *article* with its DOI provided.

Ref.	DOI	DT classification		Brief description
		WoS	Scopus	
5.2	https://doi.org/10.1016/j.dib.2018.10.142	*Article; data paper*	*Article*	Same as 5.1, but classified as a *data article.*
5.3	https://doi.org/10.1016/j.dib.2018.11.129	*Article; data paper*	*Article*	Same as 5.2.
6.1	https://doi.org/10.1007/978-1-0716-0603-2_5	*Article; book chapter*	*Book chapter*	Special document consisting of a *protocol* published in a book dedicated to operational *methods and protocols* in medicine/biology.
6.2	https://doi.org/10.1007/978-1-0716-0611-7_11	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.3	https://doi.org/10.1007/978-1-0716-0978-1_14	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.4	https://doi.org/10.1007/978-1-0716-0978-1_25	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.5	https://doi.org/10.1007/978-1-0716-0978-1_27	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.6	https://doi.org/10.1007/978-1-0716-0978-1_38	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.7	https://doi.org/10.1007/978-1-0716-0978-1_40	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.8	https://doi.org/10.1007/978-1-0716-1174-6_14	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.9	https://doi.org/10.1007/978-1-4939-7584-6_8	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.10	https://doi.org/10.1007/978-1-4939-8837-2_3	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.11	https://doi.org/10 1007/978-1-4939-8982-9_15	*Article; book chapter*	*Book chapter*	Same as 6.1.
6.12	https://doi.org/10.1007/978-1-4939-9873-9_5	*Article; book chapter*	*Book chapter*	Same as 6.1.

- Another common pairing of content-related and container-related DTs concerns contributions derived from conferences. These contributions, primarily articles, are commonly classified by WoS as *article; proceedings paper*, accounting for ~40% of the documents with dual-DT assignments (i.e., 423 out of 1,085; see Table 1). Although WoS's official definitions for *article* and *proceedings paper* are somewhat convoluted and appear to reference dual DTs only in specific cases, empirical observation shows that WoS systematically applies dual-DT assignments for *articles* published in special issues derived from conferences.
- An exception to the previous point arises for *reviews* originating from conferences. According to WoS's official definition: "*Review articles that were presented at symposium or conference will be processed as proceedings papers*" (Clarivate, 2025), which means that such papers are classified as pure *proceedings papers* without dual-DT assignments. This choice appears inconsistent with the dual-DT-assignment policy for *article; proceedings paper*. It would likely be more consistent to also allow dual-DT assignments such as *review; proceedings paper*. For example, document 2.1 in Table 5 is a *review* clearly derived from a conference, as explicitly stated in the journal special issue where it appears.

However, WoS erroneously classifies it as *article; proceedings paper* instead of *review; proceedings paper*.

- From Scopus's perspective, articles from journal special issues linked to conferences are simply classified as *articles*, effectively equating them with "pure" journal articles, which are typically subjected to a more rigorous selection process. For instance, documents 3.1 to 3.7 in Table 5 come from three different conferences: the first three, linked to CEST 2017 and NAXOS 2018, were published in special issues of the journal *Desalination and Water Treatment*. The next three, from the *Random Matrix Theory* conference in Kraków, appeared in *Acta Physica Polonica B*, a journal exclusively dedicated to conference proceedings. The final document (3.7) was published in a regular issue of the *Journal of Clinical Oncology* but had been presented at three different conferences[2]. Both WoS and Scopus correctly classify these seven documents based on their internal criteria. However, as with *book chapters*, Scopus's convention for journal special issues results in a loss of information regarding their conference origin, effectively conflating them with pure journal articles. This sort of "promotion" can impact bibliometric indicators – at the journal, researcher, or institutional level – which may not always differentiate between regular and special issues of journals (Franceschini et al., 2019).

(2) *Early-access* documents. A substantial portion of the analysed documents (i.e., 454 out of 1,085, corresponding to ~42%; see Table 1) received dual-DT assignments in WoS, where the primary DT refers to the content of the contribution, typically in a scientific journal (*article*, *correction*, *review*, or *editorial material*), and the secondary DT corresponds to the temporary designation of *early access*. This designation indicates that the contribution has been accepted and made publicly available online but has not yet appeared in its final editorial format (e.g., with volume, issue number, and definitive page numbers). The points below summarise some curious aspects observed regarding this category of documents.

- Although Scopus officially includes *article in press* among its defined DTs (Elsevier, 2025), it does not appear to use this designation in practice. As a result, documents labelled as *early access* in WoS generally do not pose classification issues for Scopus, except for occasional misclassifications between *article* and *review* – a phenomenon already observed in previous studies (Donner, 2023; Haupka et al., 2024; Zhu et al., 2024; Maisano et al., 2025).
- Returning to WoS, the manual analysis revealed that *early-access* documents are generally classified correctly, except for some misclassifications of the primary DT, particularly between *article* and *review*. Regarding the secondary *early-access* DT, ~85% of the analysed documents were found to be correctly classified.

[2] In fact, a dedicated section of the paper, named "Prior Presentation", reads: "*Presented at the 59th Annual Meeting of the American Society of Hematology, Atlanta, GA, December 9-12, 2017; the meeting of the American Society of Blood and Marrow Transplantation, Salt Lake City, UT, February 21-25, 2018; and the 16th International Umbilical Cord Blood Symposium, San Diego, CA, June 14-16, 2018*".

However, ~15% were already in their final editorial format, meaning the temporary *early-access* designation should have been removed. As noted in the previous subsection, this inaccuracy does not appear to be particularly severe. Furthermore, a follow-up check conducted in December 2024 (approximately eight months after the initial data retrieval and analysis) revealed that nearly all previous anomalies had been corrected by WoS (Franceschini et al., 2016b).

- As an anecdote, one document (document 4.1 in Table 5) highlights a rare double misclassification. This unique document – a *Christmas song* published in a scientific journal – was not only no longer *early access* but was also misclassified by WoS as *editorial material*. A more appropriate classification might have been *other*. Scopus also misclassified it too as an *article*, whereas *note* would probably have been a more suitable designation.
- Another curious observation involves *early-access* designations for articles in journal special issues or extended versions of conference contributions. Occasionally, this temporary designation seems to "overwrite" a potential secondary DT of *proceedings paper*. Once the article, initially classified as *article; early access*, is published in its final form, WoS does not appear to replace *early access* with *proceedings paper*, which would seem appropriate as it is typically associated with journal articles originating from conferences, as previously documented. For instance, consider documents 4.2 and 4.3 in Table 5. Both are extended versions of articles originally presented in conference proceedings. However, they do not carry the dual-DT assignment *article; proceedings paper* once published in their final form. It appears that the temporary *early-access* designation displaces the secondary *proceedings paper* DT, and the database fails to reinstate it after final publication. This observation warrants further investigation in future studies. While Scopus's policy of assigning only a single DT introduces potential inaccuracies, WoS's limit of a maximum of two DTs may sometimes lead to inaccuracies, as exemplified by the issue discussed above.

(3) Uncommon, specialized journal documents. A less frequent scenario in which WoS assigns dual DTs involves documents published in journals that differ from traditional contributions (*articles*, *reviews*, *letters*, etc.). Below are some of the most interesting cases observed.

- Among these less common documents, we identified forty-five so-called *data papers*, which are essentially documents containing detailed datasets that support other scientific contributions (typically journal articles) to which they are linked. According to its internal rule, WoS classifies these contributions with a dual-DT assignment: *article; data paper* (cf. the definition: "*A data paper will have a dual document type: article; data paper*" (Clarivate, 2025)). Conversely, Scopus, despite having a dedicated *data-paper* DT (Elsevier, 2025), sometimes classifies these documents simply as *articles*. Table 5 lists three examples of such documents (5.1 to 5.3). The most critical consequence of these inaccuracies is the undue "promotion" of *data papers* to the level as journal *articles*.

- Among the less common documents with dual DTs in WoS, we also found one classified as *article; retracted* and another as *article; expression of concern*. Scopus, in comparison, categorized these documents into its dedicated *retracted* and *article* categories, respectively.
- Finally, we draw attention to another uncommon category of documents: *protocols* or *methods and protocols*. These documents, which primarily detail best practices in medicine/biology, do not have a direct counterpart in the DT categories of either WoS or Scopus. Since these contributions appear almost exclusively in book series, Scopus indexed them – correctly, in our view – as *book chapters*, while WoS assigned them the dual DT *article; book chapter*. While WoS's classifications were not deemed erroneous in our analysis, it might have been more appropriate to classify these documents as *other; book chapter*. In fact, the content of these *protocols* often lacks the structure of canonical articles, making the *article* designation less fitting. Table 5 exemplifies the twelve documents (6.1 to 6.12) identified during the analysis.

Conclusions

This research focused on scientific documents with dual-DT assignments in WoS, which – based on a preliminary estimate – constitute ~4% of all indexed documents. The aim was to identify potential issues in DT classification, not only from the perspective of WoS but also Scopus. Manual analysis of a *corpus* of 1,085 documents revealed that documents with dual DTs are more prone to classification errors than those with single DTs: error rate of 4.9% versus 2.3% for WoS and 7.9% versus 2.7% for Scopus. Thus, addressing **RQ#1**, it can be concluded that these documents significantly differ from those with single DTs, as confirmed by appropriate statistical tests.

In general, three main scenarios were identified where WoS uses dual-DT classification:

1. Cases where the primary DT specifies the *content* type (e.g., *article, review, letter*), while the secondary DT specifies the *container* (e.g., *book chapter, proceedings paper*), if different from the journal container implicitly referenced by WoS for single-DT classifications.
2. Less common and specialized documents usually published in journals. These documents are relatively few and do not significantly impact overall error statistics. In cases where the specific DT is not covered by the database's predefined categories, we suggest avoiding overuse of the DT *article* and instead replacing it with a "catch-all" DT *other*. This would avoid various undue "promotions".
3. Documents temporarily assigned the *early-access* secondary DT, while awaiting their final published format. A notable number of errors stemmed from the failure to update *early-access* journal documents (*articles, reviews, letters,* etc.) in WoS after their final publication.

The analysis also revealed that some DT-classification errors may stem from inconsistencies or ambiguities in DT definitions. For instance, WoS's definition of

book chapter and Scopus's definitions of *article* exhibit certain ambiguities. Nevertheless, allowing dual-DT assignments in WoS serves as a "safety net" to avoid relatively severe errors, which Scopus sometimes encounters. For WoS, cases where both assigned DTs are incorrect (*full* errors) represent only 10 out of 1,085 documents, whereas in 87 cases, one of the assigned DTs was correct (*partial* errors). Scopus's strict single-DT assignment policy appears to be a relevant cause of its misclassifications. This limitation arises from the simple fact that DTs are not always mutually exclusive; in some cases, multiple DTs may be valid simultaneously (e.g., a *review* and a *conference paper*). Although forcing a single-DT assignment might seem like a simplification, this approach can lead to potential errors, such as undue "promotion" (e.g., from *proceedings paper* to journal *article*) or, at least, a loss of information about the documents in question. For this reason, WoS's policy of allowing dual-DT assignments seems more prudent (**RQ#2**). From a practical standpoint, it might even make sense to use up to three DTs in certain cases: one for *content*, one for the *container*, and one for an *accessory* designation (e.g., *early access*, *retracted*, etc.).

The findings of this study have practical implications for several stakeholders. For individual researchers, they may provide additional guidance for collecting and selecting documents from scientific literature through databases. For bibliometric indicator developers, this study raises awareness of potential distortions caused by DT classification errors, which have been at least preliminarily quantified here. For database managers, the comparative analysis of current DT-assignment policies could inform future improvements in DT definitions and their assignment logic.

In general, we recommend that database providers refine and clarify their DT classification guidelines to minimize ambiguities (e.g., clearly distinguishing an *article* from an *editorial material* or *proceedings paper*) and consider integrating AI-based tools to assist in the DT classification process. Automated checks – using machine learning trained on document metadata and *full texts* – could help flag inconsistent or unlikely DT assignments for human review, thus improving the overall accuracy of the databases. Additionally, relaxing the current WoS-imposed limit of two DTs to allow a third DT could be a reasonable step forward.

Furthermore, it would be useful to consider how other bibliometric databases handle DT classification to put these findings in context. For example, the *Dimensions* bibliometric platform (developed by Digital Science) combines publisher metadata and machine learning to assign DTs and links them to research grants, patents, and policy outputs (Digital Science, 2025). Open scholarly platforms like *OpenAlex* rely largely on publisher-provided metadata (via *Crossref*) for DTs, resulting in a broader but less standardized set of DTs. Studies have shown that DTs can differ considerably between providers, and what counts as a "research" document versus "non-research" can vary by database. These discrepancies underscore the absence of a universal standard for DT classification across database, which can complicate cross-database comparisons. Our findings and recommendations align with recent calls for richer and more consistent DT metadata in bibliometric data sources (Haupka et al., 2024).

The primary limitation of this research is the relatively small sample size (1,085 documents), which may hinder the generalizability of the findings. In future work, we aim to extend the sample size to provide a more comprehensive analysis.

Acknowledgments

This research was carried out under the MICS (Made in Italy, Circular and Sustainable) Extended Partnership and partially funded by the European Union Next-GenerationEU (Piano Nazionale di Ripresa e Resilienza – Missione 4, Componente 2, Investimento 1.3, D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

Clarivate (2025). Web of Science Help – Document Types, https://webofscience.help.clarivate.com/en-us/Content/document-types.html [accessed April 2025].

Digital Science (2025) Which publication and document types are available in Dimensions? https://dimensions.freshdesk.com/support/solutions/articles/23000018866-which-publication-and-document-types-are-available-in-dimensions- [accessed April 2025].

Donner, P. (2017). Document type assignment accuracy in the journal citation index data of Web of Science. *Scientometrics*, 113(1), 219-236.

Donner, P. (2023). Data inaccuracy quantification and uncertainty propagation for bibliometric indicators. *arXiv preprint*, arXiv:2303.16613

Elsevier (2025). Scopus Content Coverage Guide, https://www.elsevier.com/products/scopus/content [accessed April 2024].

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2015). Errors in DOI indexing by bibliometric databases. *Scientometrics*, 102, 2181-2186.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2016a). Empirical analysis and classification of database errors in Scopus and Web of Science. *Journal of Informetrics*, 10(4), 933-953.

Franceschini, F., Maisano, D., Mastrogiacomo, L. (2016b). Do Scopus and WoS correct "old" omitted citations? *Scientometrics*, 107, 321-335.

Franceschini, F., Galetto, M., Maisano, D. (2019) *Designing Performance Measurement Systems: Theory and Practice of Key Performance Indicators*, Springer International Publishing. Cham, Switzerland, ISBN: 978-3-030-01191-8.

García-Pérez, M.A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in psychology. *Journal of the American Society for Information Science and Technology*, 61/10: 2070–85.

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., Mayr, P. (2024). Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *arXiv preprint*, arXiv:2406.15154.

Maisano, D., Mastrogiacomo, L., Ferrara, L., Franceschini, F. (2025). A large-scale semi-automated approach for assessing document-type classification errors in bibliometric databases. *Scientometrics*, 130, 1901-1938.

Mokhnacheva, Y.V. (2023). Document Types Indexed in WoS and Scopus: Similarities, Differences, and Their Significance in the Analysis of Publication Activity. *Scientific and Technical Information Processing*, 50(1), 40-46.

Ross, S.M. (2017). *Introductory statistics*. Academic Press, London, ISBN 978-0-12-804317-2.

Yeung, A.W.K. (2019). Comparison between Scopus, Web of Science, PubMed and publishers for mislabelled review papers. *Current Science*, 116(11), 1909-1914.

Yeung, A.W.K. (2021). Document type assignment by Web of Science, Scopus, PubMed, and publishers to "Top 100" papers. *Malaysian Journal of Library & Information Science*, 26(3), 97-103.

Zhu, M., Lu, X., Chen, F., Yang, L., Shen, Z. (2024). An explorative study on document type assignment of review articles in Web of Science, Scopus and journals' websites. *Journal of Data and Information Science*, 9(1), 11-36.

https://doi.org/10.51408/issi2025_003

Influence of Regulation on Research and Technology Maturation: A Bibliometric Investigation of Research in Aftertreatment Technology

Sujit Bhattacharya[1], Sandhiya Laksmanan[2], Lata Kashyap[3]

[1]*sujit_academic@yahoo.com*, [2]*sandhiya@niscpr.res.in*, [3]*kashyaplata445@gmail.com*
CSIR-National Institute of Science Communication and Policy Research, Dr K.S.Krishnan Marg, New Delhi (India)

Abstract

Environmental risk has emerged as a primary concern globally, with air pollution being the most significant contributor to this risk. The United Nations Framework Convention on Climate Change (UNFCCC), the Paris Agreement, Net Zero emission targets, and growing concern about climate change and air pollution are pushing for rigorous emission norms. Vehicular emission has been identified as a major contributor to air pollution. Health alarms and increasing public concern have led to the development of stringent regulations for abating vehicular emissions. Efficiency in controlling vehicular emissions (Aftertreatment Systems) has emerged as one of the critical factors for the competitiveness of automobile companies. Aftertreatment systems are complex systems with various components that must be integrated to develop an effective mechanism for vehicular emission control. Scientific leads provide the key inputs for the technology development of aftertreatment systems. The influence of regulation on scientific research in an area provides a novel understanding of a demand-pull model of research, increasing regulation creating demand driven research. This aspect has not been researched to that extent and more so applying the bibliometrics approach. The present paper is centred in this direction. It examines the key regulations implemented over different periods for controlling vehicular emissions and their influence on aftertreatment research. A sophisticated bibliometric science mapping approach is applied to capture the temporal and longitudinal evolution of research in aftertreatment, covering the period from 1991 to 2023. The paper provides interesting insights into research shaped by regulations and concludes by drawing policy implications.

Introduction

Environmental risk has emerged as a primary concern globally, with air pollution being the most significant contributor to this risk. WHO has prescribed guidelines for air quality standards that provide countries with a clean air benchmark for avoiding health risks due to poor air quality. Clean air has a significant positive impact on Sustainable Development Goals; attaining clean air has emerged as one of the key targets set by different countries for attaining SDG goals. The air quality standards of countries have been set at different levels, and OECD countries have set high standards. Research over the years identified the major pollutants and their effect on health. Vehicular emission was identified as one of the major contributors to air pollution; progressively, many pollutants emitted from tailpipes were found to have severe adverse impacts on the air. This raised health alarms and increasing public concern led to stringent regulations for abating vehicular emissions. The

major effect on global automobile market was visible with the need to bring effective interventions for abating vehicular emissions. The case of Volkswagon highlights how contentious and severe implications can happen due to violation of emission standards. Volkswagon was caught by the U.S. Environmental Protection Agency (EPA) in installing a 'defeat device', a software in diesel cars and SUVs that helps circumvent EPA emissions standards for certain air pollutants. It was designed to detect when the vehicle is undergoing emissions testing and turns full emissions controls on only during the test. These vehicles were found to emit up to nine times more pollution than emissions standards allow. This also created a huge reputation damage on Volkswagon globally and also it had to pay a $2.8 billion to settle the penalty imposed. It is estimated as early as 2014 that over 70 percent of light vehicles sold worldwide are subjected to vehicle emissions standards (IIS 2014). Huge capital is being devoted by major automobile companies for developing effective post-treatments (aftertreatment technologies). Research provides the leads for technology pathways for developing exhaust-abating products. R&D efforts in the development of sophisticated aftertreatment systems is driven in response to the increasing regulation. In other words, the strict emission standards are leading to the demand for development of sophisticated aftertreatment systems; a 'demand-pull' driven by market needs of creating aftertreatment systems that complies with emission regulations (Frenkel et al., 2014).

Emission standards have become the legal requirements governing air pollutants released into the atmosphere. Emission standards set quantitative limits on the permissible amount of air pollutants released from specific sources over specific timeframes. They are generally designed to achieve air quality standards and to protect human life. Vehicular emission standards have evolved and become more stringent over the years limiting the amount of pollutants that vehicles and engines can emit. The mechanism of fuel combustion and exhaust emissions largely relies on fuel and engine design properties. Many approaches are used to reduce engine exhaust emissions into the atmosphere. An integrated approach of considering both 'internal factors' that result in better engine combustion and 'aftertreatment' technologies that can reduce already borne pollutants in the exhaust stream is needed for emission reduction from engines. The internal factors influence the combustion chamber and fuel delivery systems. An aftertreatment system is a method or device for reducing harmful exhaust emissions from internal combustion engines. In other words, it is a device that cleans exhaust gases to ensure that the engines meet emission regulations. The aftertreatment systems are used to reduce NOx, CO, PM, and BC emissions and are continually evolving to control tailpipe emission pollutants that are identified as causing environmental risks and health hazards. Balancing the factors is highly challenging as some conditions favour the reduction of these emissions, while some situations lead to an increase in the emissions (Gajbhiye et al., 2022). An amalgamation of internal factors and aftertreatment technologies in engines is a big challenge in modern engine technology.

Reducing the pollutants deriving from engines represents an interesting scientific and technological challenge. It has high commercial value as upstream research is a crucial driver of invention and innovation. As regulation standards are getting more stringent globally and public concerns are increasing, the need for more sophisticated aftertreatment systems is increasing. In particular, one of the critical factors for the competitiveness of automobile companies is their efficiency in controlling emissions. Research is happening intensively to develop advanced filters for creating more efficient aftertreatment systems. Vehicular tailpipe emissions significantly threaten human and environmental health (WHO, 2019). Incomplete combustion in the engine is one of the significant reasons for combustion pollutants, with hydrocarbons (HC), carbon monoxide (CO), and nitrogen oxide (NO_x) being the significant pollutants realized during this process. HC is responsible for soot formation and Particulate matter ($PM_{2.5}$), which negatively affect human health. CO reduces the flow of oxygen in the bloodstream and is particularly dangerous to a person with heart disease. CO_2 does not directly impact human health, but it is a 'greenhouse gas' that traps the earth's heat and contributes to the potential for global warming (Jalali et al., 2022). NO_x is the major pollutant that acts as a precursor for tropospheric ozone (O_3) formation, resulting in several allergenic diseases. The vehicular emissions thus has severe implications for adverse impact on clean air. Clean air is now closely monitored through various instruments in different countries. Clean Air Act in the USA has set national ambient air quality standards (NAAQS) primarily focussing on six major air pollutants: particulate matter (PM10 and PM2.5), ground-level ozone (O3), Nitrogen dioxide (NO20, Sulfur dioxide (SO2), Carbon monoxide (CO), Lead (Pb). Environmental Protection Agency (EPA) is the regulatory authority in the USA that monitors environmental pollution, within its ambit is seeing that automobiles manufactured domestically and imported do not violate vehicular emission. European Clean air policy calls for member states to adopt measures that as a minimum level aligns with WHO air quality standards. Other countries are also adopting stringent guidelines for Clean Air. Emission reduction measures are at the centre stage of the policy programmes in EU. Horizon Europe and Cohesion fund to support clean air initiatives by at the local, regional, and national levels.

Increasing evidence of vehicular emissions being one of the key factor of air pollution is leading to framing more stringent emission norms. Research is also highlighting newer pollutants from vehicular tail pipes that need to be controlled and limiting particle size ranges with varied dimensions from 10-300 nanometers (nm). The US standards are designed in terms of Tiers with the current being Tier 4, Europe has Euro standards with current being EURO 6/VI regulations, China has stage 6 emission standards. Standards in different countries have been primarily framed from Euro and US standards. India for example has BS standards primarily framed from Euro standards. Euro 7 exhaust emission is introducing stricter regulations for various pollutants and regulations on contaminants that have not been regulated until

now, such as nitrous oxide (N_2O). The US is implementing new rules limiting fine particulate pollutants to 9 ug/m3 (micrograms per cubic meter of air). Meeting the stringent limits requires a catalytic system with great complexity, size of units, and number of units, as well as increased fuel consumption. Overall, the common theme among these regulations is the focus on reducing emissions of harmful pollutants from vehicles to improve air quality and human health. However, there are differences in the specific pollutants targeted and the standards set. US EPA ensures stricter compliance for domestic and important automobiles with the regulation for vehicular emission control. Similar measures are visible for EU member countries. Other countries are also adopting measures in similar directions.

Objective and Research Questions

Regulation standards for vehicular emissions are becoming more stringent, expanding the scope of pollutants and particle size range with varied dimensions from 10-300 nanometers (nm) that need to be controlled. This creates concerns for automobile manufacturers as they have to demonstrate to the regulatory authorities that their vehicle meets the vehicular emission standards of that country. Automobile firms compete to establish efficient aftertreatment systems, pushing much attention to research and development. Countries compete in manufacturing and export, with the automobile sector as one of the key areas that bring economic growth, jobs and a long-term competitive edge. It directly and indirectly contributes to many related industries. The countries, particularly those having strong automobile sectors, are also devoting substantial funding to automobile research and technology development. With emission control becoming a key area, funding for developing different components for efficient aftertreatment systems has become an active area of research.

Research papers can provide a good indication of what types of research are happening in cutting-edge areas that contribute to developing technological capacity. Research papers are also published in cutting-edge areas to defend against competitor firms and prevent them from patenting. The key argument the study makes is that stringent regulations is pushing research in this area towards the development of sophisticated aftertreatment systems. We posit that aftertreatment research is shaped by regulations over different periods. The study will provide an idea of how, in a demand-driven area which is highly science-intensive, regulation motivates research.

The following research questions are posed to address the study's objectives, i.e., the influence of vehicular emission regulations on research in aftertreatment technologies.

- What pollutants are regulated in different periods, and how does it map with research activity?
- How central and developed are those research themes in different periods?

- Which are the most important topics in terms of research productivity and impact?
- What is the overall landscape of research in this area, and what it indicates?

Literature Review

Knowledge embodied in publications are important explicit outputs of R&D. Publication analysis using tools and techniques of bibliometric offers the advantage of enabling an objective, quantitative analysis of prominent, explicit outputs of R&D. Bibliometric has been applied extensively to understand the structure and dynamics of a research field, its intellectual structure, trends and evolution (Leung, Sun and, Bai, 2017). The relationship between disciplines, fields, documents, or authors can be spatially represented through bibliometric mapping (also called science mapping) (Small, 1999). New sophisticated tools and techniques of bibliometrics are helping in diving deep to understand the evolution of a research field, the emerging key areas, and other essential insights (Cartes-Velásquez and Manterola-Delgado, 2014).

A few bibliometric studies have been undertaken in vehicular emissions and subdomains of aftertreatment systems. Egan, Mohammadpour, and Salehi (2023) present a bibliometric analysis of research from ship emissions. The temporal evolution of the field was undertaken based on keywords. The two topics, 'energy efficiency' and 'emission reduction', were found to be prominently dominating from 2014 to 2020. Additionally, the increase in the term 'climate change' occurs in the same period as terms such as 'LNG' and 'emission control', showing the increasing trend towards sustainable maritime transport. This is seen again between 2021 and 2022 in terms of 'green shipping' and 'fuel sulfur content', as well as 'energy efficiency' remaining a frequently occurring term. This provides a good indication of the topics where research is taking place. Tian et al. (2018) identified trends and characteristics of carbon emissions research in the transportation sector from 1997–2016. Bibliometric analysis was based on keywords assigned to publications to identify critical topics in this field. Ruegg and Thomas (2011) examined the extent to which US Department of Energy-supported combustion engine research and how it was linked to downstream advances in vehicle engines and innovation in other industries. Huang et al. (2017) attempted to uncover the frontier research areas in selective catalytic reduction Technology (SCR). Ai et al. (2023) analysed the global research landscape and hotspots in SCR. Their study identified five major SCR research areas: catalyst, reductant, deactivation, mechanism, zeolite. Zeolite was found to be the most widely studied SCR catalyst.

Analysis of research evolution in aftertreatment systems covering combustion engines and Selective Catalytic reduction technology has been undertaken so far. There is, however, no study that has captured the overall research landscape of the aftertreatment system and the different components within it. No study in this area has examined the regulations and its implications for aftertreatment research. The

present paper is motivated by this as it attempts to see how emission regulations shape aftertreatment research.

Methodology and Data Collection

Emission control regulation for vehicular emission was examined from key influential sources primarily from European and US regulations. The evolution of regulations in these two countries has influenced regulations in other countries. It thus provides a good measure to capture the pollutants identified and specific attributes for control over different periods. The study uses a highly sophisticated science mapping tool, SciMAT, to capture the themes or topics in specific delineated periods (temporal analysis) and track the evolution of research fields through consecutive periods (longitudinal study). This software applies the bibliometric technique of co-word analysis, helps to construct a co-word network through co-word analysis and applies a clustering algorithm to identify the research themes (Coulters, Monarch & Konda, 1998). Two dimensions, namely 'centrality' and 'density' characterise each topic (Callon, Courtial, and Laville, 1991). Centrality measures the external interaction among each network and can be understood as the relevance value of the topic. The internal cohesion of the network is measured by density; it can be interpreted as a measure of the theme's development. Based on centrality and density, SciMAT allows a research field to be represented through a Strategic diagram and thematic network. A strategic diagram is a temporal analysis of one selected period. A four-dimension quadrant can represent the critical characterization, see Figure 1.

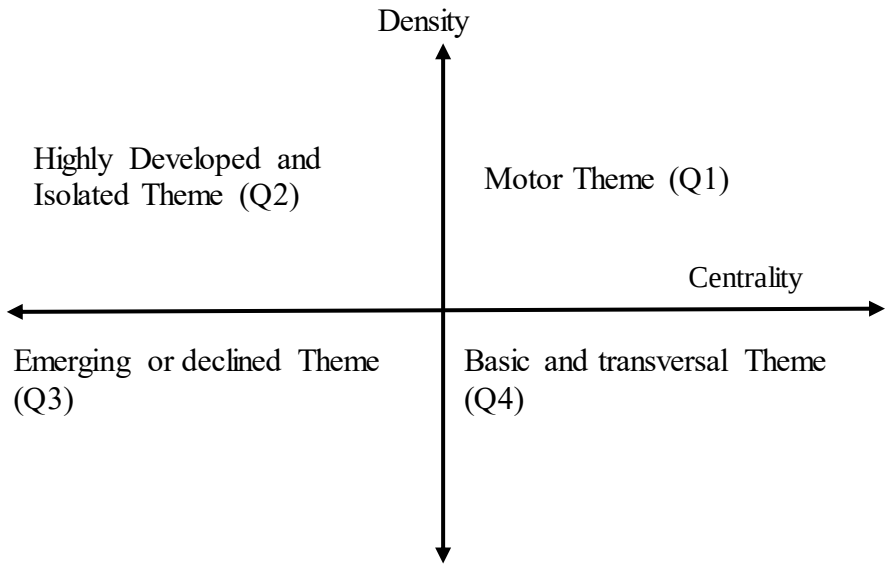


Figure 1. A Stylised Representation of Strategic diagram.

These quadrants are defined as follows (Cobo et al., 2012): The upper-right quadrant (Q1) is identified as "Motor themes", characterised by internal solid ties among the sub-topics (high density) and also connecting activity with other themes (high degree of centrality). The upper-left quadrant (Q2) hosts topics with strong internal ties but weak external links. They stand as specialized themes on the area's periphery. The lower-left quadrant (Q3) includes themes in which sub-topics are not connecting to each other (low density), and also, they are not connecting to different themes (low centrality). They can be emerging or declining themes. The lower-right quadrant (Q4) has themes in which sub-topics are loosely connected (low density) but are well connected to different themes (high centrality). A longitudinal analysis leads to a thematic area, a set of themes that have evolved across different sub-periods. The other themes are nodes in a network. Themes are connected to a network through edges. This helps to create a thematic area based on longitudinal analysis. Various indexes can be applied to normalize the data and provide weightage. A theme could belong to a different thematic area or not come from any. The coherence and diversity that lead to the formation of an area can be discerned through this analysis. It is more valuable if quantity and quality performance indicators are applied to understand the theme development and evolution. Quantity measured through publication count is a proxy for research intensity, and quality is measured through citations and citation-based indicators that act as a proxy for research influence and impact. The present paper used this comprehensive approach: a strategic diagram and thematic network constructed based on publication count, total citations, and h-index (Hirsch, 2005) to capture the impact of themes and thematic areas.

Data set

The metadata for the study was based on papers downloaded from the Web of Science (WoS). Core components of aftertreatment technology on which all the technologies have been primarily based were identified based on a literature review of aftertreatment technology and consultation with subject experts. This helped to develop the search string to extract records for the study. Selective Catalytic Reduction (SCR) and the lean NOx trap (LNT) are the representative technologies devoted to reducing NOx under lean-burn operation conditions. At the same time, soot removal is mainly performed by Diesel Particulate Filters (DPF). Various new combined technologies have been introduced for NO_x removal (i.e., LNT–SCR) and the simultaneous removal of NO_x and soot, like SCR-on-filter (SCRoF), in series LNT/DPF and SCR/DPF and LNT/DPF and SCR/DPF hybrid systems (Marinovic et al., 2022). Diesel Oxidation Catalysts (DOC) are comprehensively preferred emission control systems for heavy-duty and light-duty vehicles in many countries such as Europe, the USA, and Japan. Gasoline particulate filters (GPFs) are emerging as helpful after-treatment filters for meeting the limits of ultrafine particles in gasoline/petrol vehicles. Highly efficient exhaust filters capture most of the vehicle's particulate emissions, including excellent particles, preventing them from leaving

the exhaust pipe or possibly entering the atmosphere. These terms were introduced to capture the records for the study. The final search string that had the slightest noise and provided the most relevant records was (("SCR", or "selective catalytic reduction" or "DPF" or "Diesel Particulate Filter" or "EGC" or "Exhaust Gas Circulation", or "DOC", or "Diesel Oxidation Catalyst" or "LNT" or "Lean NOx trap" or "GPF" or "Gasoline particulate Filters") and ("aftertreatment*")). This search string was applied to the Topic Search, which includes the topic, abstract, authors' keywords, and keyword plus (indexed words assigned to papers by WoS). The study covered the period from 1991, when the first publication on this topic appeared in this database, till 2023. Keywords were the most critical analytical unit for the study as they are seen as a signal to the fundamental concepts/topics of the research paper (Bhattacharya and Basu, 1998). The author, journal, and keywords given by the WoS database were taken for each document to have a more exhaustive corpus.

The extracted data was pre-processed using SciMAT to remove duplicate and misspelt terms. Corresponding full names replaced abbreviations with a mapping table, e.g., SCR by Selective Catalysts Reduction and Particulate Matter by PM. This helped create a consolidated group of keywords representing the same theme. The research activity in this field has been prominently visible since 1995. Hence, the period taken was from 1995-2023 for in-depth analysis. This period was delineated into six phases: 1995- 2007, 2008-2011, 2012–2015, 2016-2019, 2020–2023, to capture the diachronic changes of the field's evolution more effectively. Also during the identified periods, regulations were introduced that had major impact on vehicular emissions.

For each phase, themes were created through a strategic diagram, and a performance table was extracted. The co-occurrence keyword was normalised using the Salton index before undertaking the co-word analysis to create a thematic network. Salton's cosine formula for normalisation was taken as it is more effective in capturing links between high and low-cited papers (Hamers et al., 1989). Area detection was done based on a thematic analysis of each period. The Inclusion Index detected the conceptual nodes (nexus) between research themes of different periods. The inclusion index is the overlap measure (e.g., Jones and Furnas, 1987; Rorvig, 1999; Salton and McGill, 1983). The inclusion index has been chosen because the weight of the thematic nexus is a good measure of the overlapping between themes. Based on the identified search string, highly cited papers were also identified for each period. This helped to provide further insights into the science maps.

Results

The study identified 802 publications in aftertreatment research indexed in WoS till 2023, with the first publication appearing in 1991. This includes 614 Articles (76% approx.), 212 proceeding papers (26% approx.), and 33 (4% approx.) review articles. Conferences in a technology-driven field bring in diverse stakeholders from academia, industry and government and primarily draw attention to how the field is

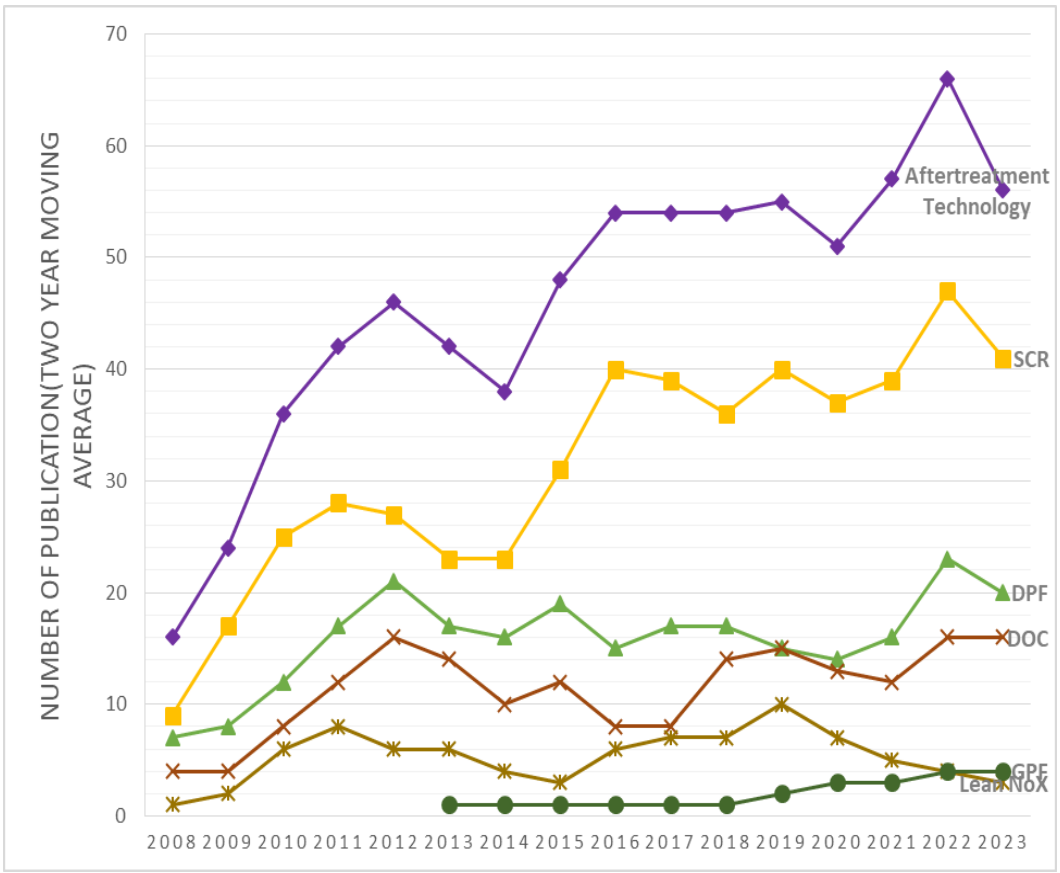


Figure 2. Research Publications in Aftertreatment and its Key Subcomponents from 2007-2023 based on two-year Moving Average.

evolving and future possibilities, among others. The conference papers indexed by WoS are also an indication of recognition of the intellectual contribution of the conference. Four key conferences, namely ASME (The American Society of Mechanical Engineers), the Internal Combustion Engine Division Fall Technical Conference, the American Control Conference, and the International Congress on Catalysis and Automotive Pollution Control, stand out as a dominant influence. All these conferences are held in the US, showing that it is the key location point for

meeting varied stakeholders such as researchers, industry, and policymakers to address the current challenges in the field.

Figure 2 highlights the publication trend overall and in key sub-domains from 2007 onwards based on a two-year moving average. The earlier period is not represented as there were, on average, three publications per year from 1991, totalling 74 publications till 2006. The overall publication trend shows increasing intensity in some key research areas/topics in the aftertreatment system. Countries have largely patterned their emission policies on European regulations making it the most widely followed emission regulation globally (ICCT_Euro 6-VI-briefing-Jun2016). The research trends thus unsurprisingly closely match with the strict regulations being introduced in different phases over the years in European Union i.e. Euro standards. The influence of US Regulations, i.e. Tier regulations, also has a strong influence as the adoption of emission standards for the road transport sector in the two leading global markets (Europe and North America) has led to the global proliferation of emission-regulated vehicles through exports (Crippa, M. et al. 2016). Development and deployment of more advanced systems and components are needed to meet the emission norms. High research activity is visible in SCR (Selective Catalytic Reduction), which is unsurprising as it is a critical component of an aftertreatment system in diesel engines. It is primarily used in reducing tailpipe emissions of nitrogen oxides (NOx) and involves several components packaged together with other parts of the emissions control system. DPF (Diesel Particulate Filter) helps collect and oxidize carbon to remove particulate matter (PM). There is a continuous demand to develop high-end DPF filters. Diesel Oxidation Catalyst (DOC) aids in this process with continuous demand to create advanced catalysts, making this an active area of research. Research activity in the Gasoline Particulate Filter (GPF) is from 2014. Strict particulate matter (PM) emission limits for gasoline engines, including particle number (PN) limits, became regulated from Euro 6 introduced in 2014. The GPF function for gasoline engines is similar to what DPF does for diesel engines. The research trends are seen in Lean NOX Traps (LNTs) from 2013. LNTs technology is used in emission control to reduce nitrogen oxides (NOx) emissions. NOx limits for vehicular emissions was introduced in Euro 5/V in 2009. However the Euro 6/VI standards introduced in 2014 significantly tightened NOx limits for diesel vehicles. Requiring maximum emission limit of no more than 80mg/km, essentially forced manufacturers to implement "lean NOx trap" technology to meet this strict regulation. Thus research trends can be observed driven by emission regulations.

The USA, Germany, and China mainly drive the publications in this area. Together, they account for 58% of the overall publications in this field. These three countries are among the top automobile manufacturing countries. China has newly emerged as one of the key players in automobile manufacturing. These countries have high stakes in emission research as it provides them with key leads for new pathways for technology development. The United States leads research on Aftertreatment

technology with 259 documents (32%). The country has implemented various measures to combat air pollution. Many of the prolific authors are also from the USA. Along with research institutions and universities, firms are also involved in research in the USA, as reflected in the publication outputs. The three firms most actively involved are from the USA, namely Ford Motors and General Motors, with 19 publications each, and Cummins (which specializes in diesel and alternative fuel engines and generators, and aftertreatment systems) with 18 publications. The other dominant countries publishing in this area are Germany and China; incidentally, both of them have 104 publications.

University-industry linkages: University-industry linkages are not strong if examined through joint publications. There may be other types of linkages that may not be captured through research papers. Cummins is most actively involved in joint publications with the university. One of Cummins' interesting linkages is with the University of Virginia. This linkage is due to the author W.S. Epling, also one of the most highly cited authors. He worked in the national laboratory (Pacific Northwest National Lab), a catalyst manufacturing company (EmeraChem), and Cummins Inc. He then again shifted to academia: the University of Waterloo, then to the University of Houston, and now the University of Virginia. His research area is on diesel engine emissions reduction, catalyst degradation, and how catalyst surfaces change with reaction conditions.

Research in this area is highly interdisciplinary: The journals in which papers are published exhibit a power law distribution, with a few journals accounting for the majority of documents. The extended tail distribution shows the scattering of papers across related fields and in multidisciplinary journals. This shows the field is highly distributed across various domains. This is an indication that aftertreatment technologies need to draw from many fields of research. The most prolific publication sources are the 'International Journal of Engine Research', 'Fuel' and 'Applied Catalysis B Environmental and Environmental Science Technology' with 42, 40 and 36 research papers contributing to 14% of overall publication in this area publications.

Temporal Analysis of Aftertreatment Technology Research

Temporal analysis was undertaken to capture the research activity in key periods when emission regulations were implemented. The main themes and sub-topics of the research activity in the different periods are represented through the Strategic diagram in Figure 2. Table 1 provides the performance analysis of each theme based on citation-based impact on two indicators: total citation count and *h*-index. Closely examining the strategic diagram, i.e. Figure 3, and the performance indicators in Table 1 can provide essential insights into the identified research themes. Each identified theme within a period can also be examined in detail by identifying sub-topics that comprise the theme. The research activity in a period, overall and in granular levels were examined in the context of emission regulations.

First period (1995-2007)

The key regulations in Europe during this period were: Euro 2 adopted in 1996 placed more stringent limits for CO and HCs+NOx (this was first introduced in Euro 1 in 1991). Euro 3 adopted in 2000 introduced HCs and NOx for regulation. Euro 5 adopted in 2005, created more stringent limits for CO, HCs and NOx. Progressive Euro regulations set stringent limits for pollutants, and more pollutants came under the ambit of regulations. This called for development of more sophisticated aftertreatment systems. For example, for gasoline engines, the Euro I limit for CO was 2.72 g/Km, Euro 3, it was 2.3 g/Km and in Euro 4 limit became more stringent at 1g/km. HC was introduced in Euro 3, setting a limit of 0.2 g/Km for subsequent Euro stages set at 0.1 g/km. Particulate Matter (PM) limit and particle number (PN) emissions were introduced only at Euro 5 and Euro 6 levels for gasoline engines. However, the PM was introduced from Euro 1 and PN from Euro 5b (in the year 2011) for Diesel engines. During this period, US introduced Tier 1, Tier II and Tier III emission regulations were introduced: Tier I in 1994, Tier II in 2000 and Tier III in 2006. Tier I standard called for light-duty vehicles to regulate CO, NOx, Particulate Matter (PM), Formaldehyde (HCHO), Non-methane organic gases (NMOG) and hydrocarbons (NMHC) emissions. For diesel engines, NOx was regulated.

Research activity could be seen pivoting in two broad research themes: 'Aftertreatment System' and 'Performance' during this period (Figure 3 and Table 1). The aftertreatment system was the most prominent in terms of research intensity and influence being in the Q1 quadrant. It is an overarching theme with 52 research papers, papers receiving good impact, and 5588 citations with an *h* index of 30. Aftertreatment research was distributed under the following major sub-topics: GPF, DPF, Particulate Matter, Vanadium, Diesel Particulate Matter, Nucleation, SCR, Biofuel, System identification, Reaction, Mathematical modelling, and Optimization. Each of the sub-topics shows the research activity in the aftertreatment system that were prominent during this period. Performance was in the Q4 quadrant, characterizing it as a primary and transversal theme. This research activity is in the following sub-topics: Hydrocarbons, Aerosol, Hydrogen, Platinum-Catalyst, and Mass Emissions. Storage/reduction catalysts and diesel particulate emission control are among the most cited articles consistent with themes. This can be partially attributed to the peak year of car manufacturing in the US around 1990. Furthermore, this is when the emission limits for NOx have been introduced, and research has primarily focused on developing catalysts for NO_x reduction.

Second period (2008-2011)

This period saw the introduction of two Euro standards: Euro 5 in 2009 and Euro 6 in 2014. More strict controls were adopted by the European Union and the US for checking that automobiles do not violate emission norms. Different countries also introduced more strict controls on emission regulations. The research activity

provides interesting insight into how research was shaped by regulations (Figure 3 and Table 1). Two Motor themes can be identified in this period: 'Air Pollution' and 'Sensors'. Research activity under the 'Air Pollution' theme is in Catalytic converter, Mathematic modelling, SCR, DPF, Fuel, Engine, EGR, Heavy duty Vehicle, Platinum catalytic, Mechanism, Impact, Optimization, Thermal efficiency, etc. This theme dominates research intensity and impact: 194 documents with 9288 citations and an *h* index 54. Many sub-topics of Aftertreatment are covered within this theme. A Sensor is installed in the exhaust gas system to recognise nitrogen oxides in the exhaust gas flow. The NOx sensor is an essential component in the aftertreatment system to reduce NOx. Sensors are often excellent indicators of pending repairs and maintenance and will point out any issues the aftertreatment system may face. This is also reflected in the thematic mapping, as this theme is centrally connected to the whole network. It also has a high citation influence of 124 despite the low research intensity of only four papers. 'Number emissions' is in the Q2 quadrant, implying strong internal cohesion within its sub-topics but weaker connections with other themes. 'Hybrid vehicle' is in the Q3 quadrant, which means it is an emerging area of research. The core functional elements of the aftertreatment system are 'Urea' and 'active catalytic'. They are in the Q4 quadrant, i.e. under the primary and general themes. This is unsurprising as they are a significant component of an aftertreatment system. The formation of ammonia from urea in heavy-duty vehicles is a precursor of secondary organic aerosols. Given this, limits on ammonia emissions are also imposed for heavy-duty vehicles and those for other pollutants in Euro standards. Platinum Catalysts in Simulated Diesel Exhaust, NH_3-SCR reactions over a Cu-zeolite, and a Fe-zeolite catalyst were the key topics in the highly cited papers. This is consistent with the critical sub-topics identified. The SCR catalysts research was emerging during this period as each catalytic converter had the issue of emitting other pollutants when one was controlled. All the SCR systems researched in this period were on a pilot scale

Third period (2012-2015)

The Euro 6 norms were strongly implemented during this period. Euro 6 emission standard specifically sets the particle number (PN) limits set for gasoline direct injection (GDI) vehicles. PN are complex mixtures of volatile and non-volatile materials containing soot, organic carbon and hydrocarbons, which required manufacturers to significantly reduce the number of emitted particles. Particulate Matter (PM) emissions from gasoline direct injection (GDI) engines, particularly Particle Number (PN) emissions were found to have adverse effects of ultrafine PM emissions on human health and other environmental concerns. Euro 6 emission standards have been introduced in Europe (and similarly in China) to limit PN emissions from GDI engines. This is prompting the development and implementation of 'gasoline particulate filter 'GPFs' an aftertreatment device to meet these stringent standards. It is particularly used in direct injection gasoline

engines, to capture and remove fine particulate matter from the exhaust gas, essentially functioning like a diesel particulate filter (DPF) but designed specifically for gasoline vehicles. It is considered a key component in modern emission control systems to meet stricter emission standards. Traps tiny soot particles present in exhaust gases from gasoline engines, preventing them from being released into the environment.

It is important to see how research shaped during this period with the advent of these stringent regulation. Research activity pivoted on ten research themes during this period (Figure 3 and Table 1). Observing the 'number of emissions' moving from the Q2 to the Q1 quadrant is interesting. In spite of only 14 papers, their influence is high as they account for 366 citations and have an index of 10. However, the dominating theme in this period is the 'Aftertreatment System' being in the Q1 quadrant. There are 304 publications under his theme with citations of 10858 and *an* impact of 55. This shows its overarching influence in terms of productivity and impact. The theme 'Vanadium' emerges as a motor theme (Q1) and exhibits good influence with 206 citations. It is interesting to see from the theme analysis of each theme that this was part of the Sensor theme in the earlier period. 'Methane' and 'Pressure' are in the Q4 quadrant (basic and transversal), with a high citation score of 82 and 90, respectively. 'Microwave technology' is the Q2 quadrant, which means well-developed interlinks (high density), but that research does not connect widely to the field, i.e. has low centrality. It has two documents and 60 citations, which shows the high reception of its papers in the community. Similar to Vanadium, this theme was part of Sensor earlier. Fast and standardized microwave heating allows delicate control of catalyst properties. Microwave-synthesized catalysts generally perform better than conventional catalysts. Highly cited papers were Diesel-engine vehicles, exhaust aftertreatment systems, low-temperature combustion, and HCCI engines with high load limits. In this period, the toxicity of emissions from diesel engines was realized, and designing catalysts for controlling emissions from diesel engines was the focus of many researchers.

Fourth period (2016-2019)

Euro 6 standards were further qualified at granular levels during this period, Euro 6a to Euro 6d. Essentially they defined with more clarity the specific characteristics of the pollutants and their emission standard limits. During this period, the Aftertreatment technology research field pivoted on eleven themes (Figure 3 and Table 1). The many themes and their position across different quadrants show the field is getting more dynamic. Researchers are working in a diversity of sub-topics. Aftertreatment technologies primarily attempt to address the challenges of air pollution. Thus, the visibility of 'Air Pollution' as a motor theme (Q1 quadrant) with a high research intensity of 382 publications is not surprising. Air pollution research in aftertreatment is widely spread across many sub-topics, as identified in the 2008-11 period. Researchers are also actively citing research on this theme, as papers have

attracted 9124 citations with a high *h*-index of 46. Other research themes that are motor themes are 'Particles', 'Sulfur Oxides', and 'On-board-diagnosis'. On-board diagnostics (OBD) a self-diagnostic system built into a system designed to ensure a vehicle is operating within emission standards for emission regulation. In the US, OBD became mandatory in 1996 for all light duty vehicles. In EU became mandatory for gasoline vehicles in 2001 and for diesel vehicles in 2003 and was first introduced in China in 2008. OBD being in Q1 quadrant shows this has become now one of the key prominent area of research now. 'Propane-oxidation' is in the basic and transversal quadrant. 'Pressure', 'Water', and 'Computational Fluid dynamic' are in the Q3 quadrant, signifying emerging research areas. 'Light-duty vehicles' and 'Energy efficiency' are in the Q2 quadrant. This is a crucial period as many developing countries like India leapfrogged from BS-IV to BS-VI norms, where aftertreatment technology compliance is the primary requisite. Highly cited papers covered the topics of Aerosol Formation from Gasoline and Diesel Motor Vehicle Emissions and Fuel Reforming and Copper catalysts. This was when the well-known connectivity between the oxidation of fuels and their contribution to air pollution was considered in real-time. The stringent emission standards were implemented, which urged automobile manufacturers to design advanced aftertreatment technologies to comply with the emission standards.

Fifth period (2020-2023)

US is introducing Tier 4 standards in 2024 with more stringent standards. Tier 4 standard calls for 90 percent reduction in PM and NOx emissions compared to Tier 1-3 standards. It also calls for ultra-low sulfur diesel fuel with a sulfur content of less than 15 parts per million (ppm). Euro 7 norm is to be introduced in EU with stricter limits on tailpipe emissions, more stringent testing requirements, and new standards for battery performance. It will also call for regulation of nitrous oxide (N_2O), this has not been regulated earlier. Having very effective control of vehicular emission has become a competitive edge for automobiles. The new regulatory norms to be introduced have already been published giving opportunity for vehicle manufacturers to be prepared for new aftertreatment systems that can adhere to these new norms. The implementation of Euro norms became more widespread with emerging economies like India also implementing norms adhering to the broad provisions of Euro 6. Most of the countries during this period had minimum EURO 4 norms. Manufacturers are also going for voluntary standards such as Blue Sky Standards that sets higher stringent limits then the existing standards. The new regulations is expected to further push investment in R&D in aftertreatment. Research intensity and areas of research scope is also expected to increase with the implemented and forthcoming new regulations. The intensity and broad spread of research during this period shows the influence of increasing regulation. This is similar to the previous period of 2016-19 when regulations became more stringent. The research during this period is seen to be distributed across 11 themes (Figure 3

and Table 1). Like the previous period, the themes are spread across the four quadrants. Each theme also has many sub-topics with which research is happening. The presence of 'Air Pollution' as a motor theme again asserts the importance of connecting research in this field to this. One can discern from the thematic map of this theme that it covers the following key domains: SCR, LNT, Aftertreatment Systems, Diesel Oxidation Catalyst, DPF, Fuel, hydrocarbons, PM, and Catalytic Converter. 'Tailpipe' is in the Q4, i.e. under basic and transversal themes. Only two research papers are receiving a high impact of 40 citations. The major challenge for aftertreatment systems is related to controlling tailpipe emissions. Surprisingly, this has not been used as a prominent keyword. *Plausibly, the broader term 'aftertreatment' or 'air pollution' is used instead of this.* 'Methane', 'Environment', 'Deposition', and 'Radio-frequency' are in the motor theme, indicating higher influence in the field. 'Air pollution' comes in the motor theme in the previous periods, showing its pervasive influence in the field. This theme is covered under the aftertreatment system in two periods. 'EGR' is in Q2 quadrant. The themes 'Marine diesel engine' and 'Real driving emission', 'Mass transfer', and 'Particulate Processes' are consolidated with emerging or declining themes. Highly cited papers in this period are regulations, current status, effects, and reduction strategies of emissions for marine diesel engines and Copper Active Sites in Zeolites by Ammonia and Plasma-Driven Nitrogen Oxidation and Catalytic Reduction.

The Figure 3 below and Table 2 shows the temporal research activity of the different periods.

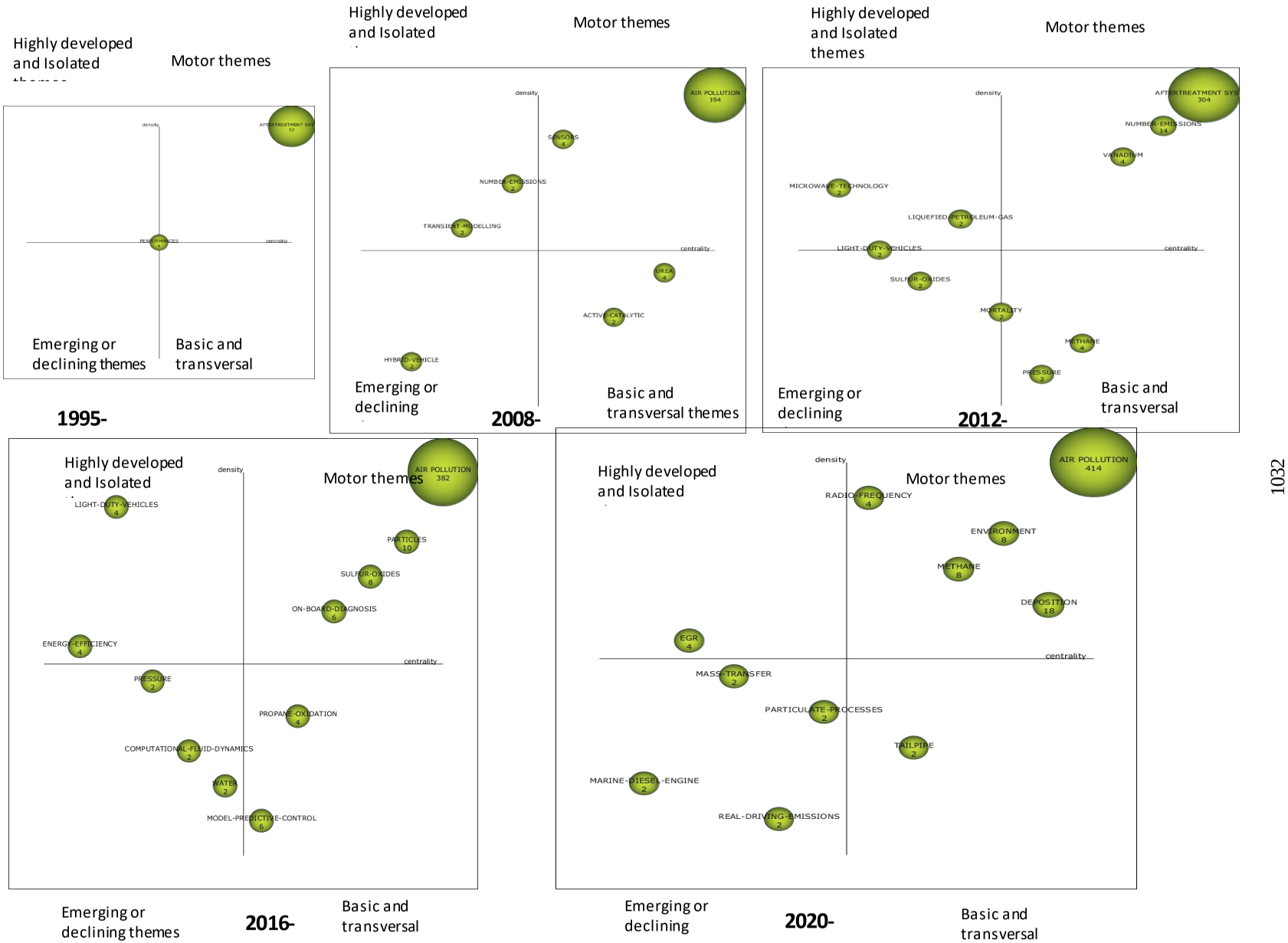


Figure 3. Strategic diagrams for the different periods in

Table 1. Performance Measurement for Different Periods.

Periods 1995-2007				Periods 2008-2011				Periods 2012-2015			
Theme	**Quadrant**	**Publications**	**Citations (*h*-Index)**	**Theme**	**Quadrant**	**Publications**	**Citations (*h*-Index)**	**Theme**	**Quadrant**	**Publications**	**Citations (*h*-Index)**
Aftertreatment System	Q1	52	5588 (30)	Active Catalytic	Q4	2	42 (2)	Aftertreatment System	Q1	304	10858 (55)
Performances	Q4	4	442 (4)	Air Pollution	Q1	194	9288 (54)	Light-Duty-Vehicles	Q2	2	366 (2)
				Hybrid Vehicle	Q3	2	0 ( )	Liquefied-Petroleum-Gas	Q2	2	60 (2)
				Number Emissions	Q2	2	62 (2)	Methane	Q4	4	82 (4)
				Sensors	Q1	4	124 (4)	Microwave - Technology	Q2	2	60 (2)
				Transient-Modelling	Q2	2	32 (2)	Mortality	Q4	2	86 (2)
				Urea	Q4	4	298 (4)	Number-Emissions	Q1	14	366 (10)
								Pressure	Q4	2	90 (2)
								Sulfur-Oxides	Q3	2	156 (2)
								Vanadium	Q1	4	206 (4)

Periods 2016-2019				Periods 2020-2023			
Theme	**Quadrant**	**Publications**	**Citations (*h*-Index)**	**Theme**	**Quadrant**	**Publications**	**Citations (*h*-Index)**
Air Pollution	Q1	382	9124 (46)	Air Pollution	Q1	414	9343 (29)
Computational-Fluid-Dynamics	Q3	2	18 (2)	Deposition	Q1	18	204 (8)
Energy-Efficiency	Q2	4	30 (4)	EGR	Q2	4	6 (2)
Light-Duty-Vehicles	Q2	4	296 (4)	Environment	Q1	8	32 (2)
Model-Predictive-Control	Q4	6	60 (6)	Marine Diesel Engine	Q3	2	30 (2)
On-Board-Diagnosis	Q1	6	38 (4)	Mass Transfer	Q3	2	6 (2)
Particles	Q1	10	230 (8)	Methane	Q1	8	68 (4)
Pressure	Q3	2	10 (2)	Particulate Process	Q3	2	8 (2)
Propane Oxidation	Q4	4	136 (4)	Radio-Frequency	Q1	4	30 (2)
Sulfur-Oxides	Q1	8	196 (6)	Real Driving Emissions	Q3	2	32 (2)
Water	Q3	2	74 (2)	Tailpipe	Q4	2	10 (2)

The strategic diagram in Figure 3 identified key themes over different temporal periods positioned under the 4 quadrants. The overall positioning of the themes and new themes emerging in different time periods is presented in Table 2. The color codes connects to the thematic evolution of the themes leading to research areas represented in Figure 4.

Table 2. Key Research Themes in Different Periods: 1995-2023.

Themes	1995-2007	2008-2011	2012-2015	2016-2019	2020-2023
Aftertreatment System	Q1		Q1		
Performances	Q4				
Active Catalytic		Q4			
Air Pollution		Q1		Q1	Q1
Hybrid Vehicle		Q3			
Number Emissions		Q2	Q1		
Sensors		Q1			
Transient- Modelling		Q2			
Urea		Q4			
Light-Duty-Vehicles			Q2	Q2	
Liquefied-Petroleum-Gas			Q2		
Methane			Q4		
Microwave-Technology			Q2		
Mortality			Q4		
Pressure			Q4	Q3	
Sulfur-Oxides			Q3		
Vanadium			Q1		
Computational-Fluid-Dynamics				Q3	
Energy-Efficiency				Q2	
Model-Predictive-Control				Q4	
On-Board- Diagnosis				Q1	
Particles				Q1	
Propane Oxidation				Q4	
Sulfur-Oxides				Q1	
Water				Q3	
Deposition					Q1
EGR					Q2
Environment					Q1
Marine Diesel Engine					Q3
Mass Transfer					Q3
Methane					Q1
Particulate Process					Q3
Radio-Frequency					Q1
Real Driving Emissions					Q3
Tailpipe					Q4

Structural analysis of the evolution of the Aftertreatment technology scientific field

A thematic area is a set of themes that have evolved across different periods. Figure 4 provides the visualisation of the analytical analysis of the themes detected in each period. The clusters of themes about the same thematic area are identified in the figure through different colours. A solid line means that both themes share one of their central keywords, and a dotted line means that both themes share some peripheral keyword (Cobo et al., 2011a). The sphere size represents the number of documents belonging to each theme. The solid lines show that the linked themes lie in the same domain. The dotted lines are connected with themes, not from the relevant domain. The thickness of the edges is proportional to the inclusion index.

Five thematic areas were identified by examining Figure 4 and Performance measures: aftertreatment *Systems*, *Sensors*, *Number Emissions*, *Active Catalytic*, and *Light-duty vehicles*. An area is formed primarily by linking themes from different periods. This shows research cohesion in this field, with only a few topics seen as isolated. In other words, research domains have evolved over various periods, with new topics included under a theme. This is not surprising as emission regulations have also evolved in this way. Air pollution and aftertreatment systems are overlapping themes with strong connectivity with each other through their central keywords. The dominance of sub-topics has sometimes resulted in the distinction between them as air pollution or aftertreatment systems. In all the periods, it is Quadrant Q1, signifying that it was the central theme with internal solid cohesion among the sub-topics and a strong connection with other themes.

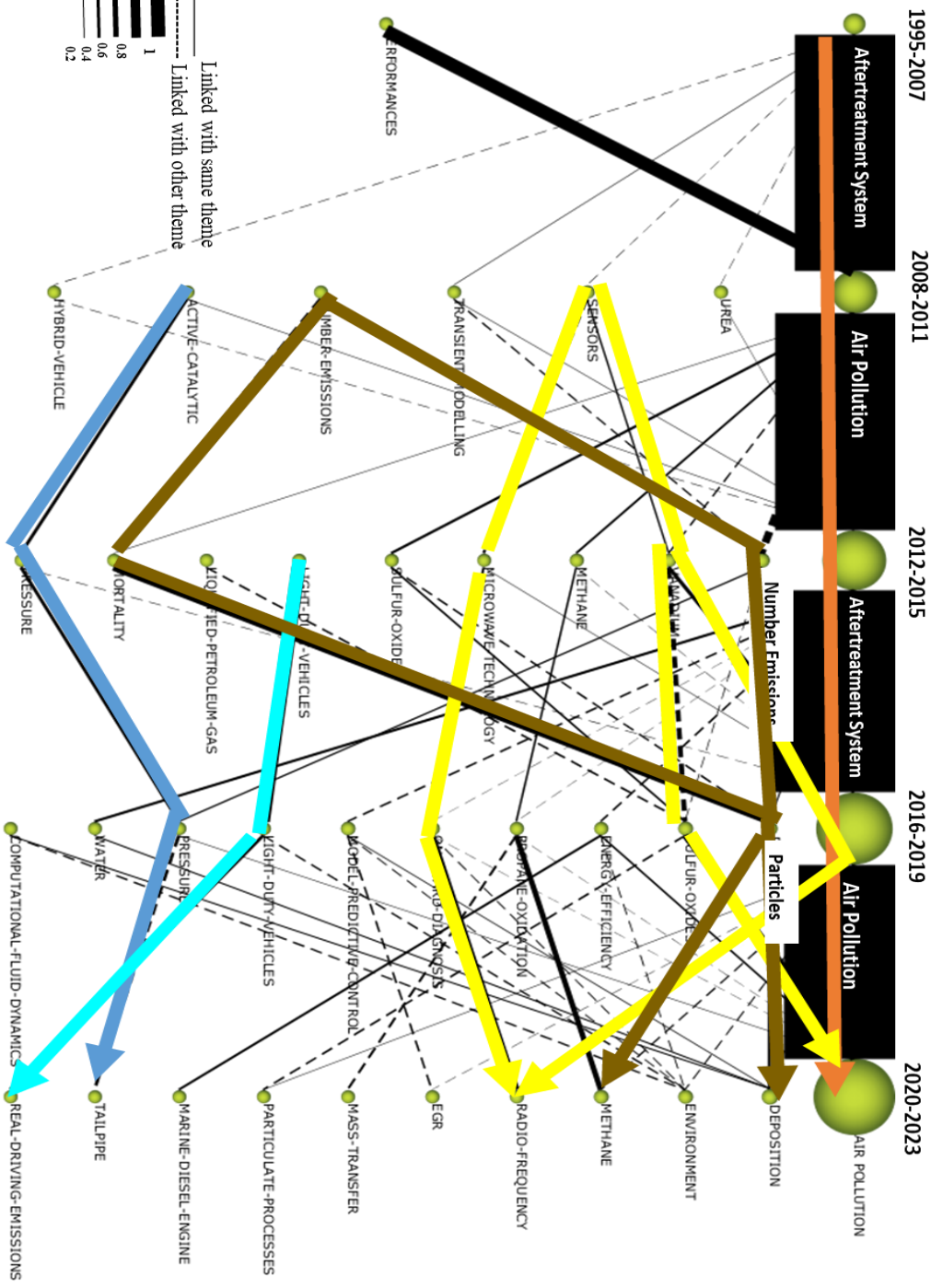


Figure 4. Thematic Evolution of Aftertreatment Technology.

The thematic area *Sensor* has evolved over different periods; research in different sub-topics related to this field has also received high citations. This theme is visible from the period 2008-2011. In the following period 92012-2015), it evolved in two thematic areas: Vanadium and Microwave technology. The theme Vanadium is in the motor theme (Q1), which shows its strong influence on research activity. Vanadium based catalysts are used in SCR processes to reduce NOx emissions. This catalyst is effective at reducing NOx at high temperatures 350-400 degree Celsius and is helping in addressing new stricter emission norms of NOx. Microwave technology is in the Q2 theme, implying strong cohesion within its different sub-topics. Microwave technology evolved from 2016 to 2019 as Onboard diagnostics (OBD). OBD is an electronic vehicle system that monitors the emissions system and critical engine components. It can usually detect a malfunction or deterioration in these components before the driver becomes aware of the problem. The regulations that called for implementing OBD and research activity happening in this domain was examined during temporal analysis. The Vanadium theme in this period is split into air pollution and sulfur oxides. Air pollution, Sulfur oxides, and OBD are in motor themes (Q1) in this period. From 2020 to 2023, Air pollution and OBD merged and evolved as Radiofrequency, and it focused on the impact on the environment. Air Pollution and Radiofrequency are in Q1 in this period. The theme particle *Number Emissions* in real-time emissions emerged in 2008-2011. It is in the Q2 quadrant, showing strong internal ties but weak external links. From 2012 to 2015, it evolved into Mortality and Number of Emission, which are in Q4 and Q1, respectively. In the next period, it evolved as Particles implies in Q1, showing the field's importance. The health effects of particulates and other vehicle emissions were analyzed in this period. From 2020 to 2023, deposition and methane evolved, which are in Q1, and show that both themes are well developed in the related field. Like other themes, the thematic area Active Catalytic shows an interesting evolutionary trend. In 2008-2011, it was in Q4 quadrant. From 2012 to 2015, it evolved as pressure in the Q4 quadrant. In the next period, the theme was in the Q3 quadrant. In the last period, it evolved as a tailpipe in Q4. The tailpipe is the last piece of the exhaust system that directs the exhaust gases out and away from the vehicle. Active catalysts are a core component of aftertreatment technologies. The period-wise thematic delineation has shown how the sub-topics have evolved in this theme. The theme Light duty vehicle evolved from 2012-2015, which is in Q2 shows the strong internal ties but weak external links. This shows that this are well-developed and isolated themes in this area. In 2016-2019, it became again in the same quadrant, as shown in Figure 4. In the last period, it evolved as a Real driving emission (RDE) in the Q3 quadrant.

This quadrant can be for emerging or declining areas. It reflects an emerging area as RDE norms are now incorporated into new regulations. It primarily measures particulate matter and nitrogen oxide emission values in real traffic. Europe was the first region to introduce RDE norms, and many other countries have adopted the European regulatory emission norms. For example, RDE norms took effect in India on April 1, 2023.

Discussion and Conclusion

Regulations are becoming essential in driving innovations and technology development. There is an increasing need to study how regulations affect scientific research as newly emerging technologies are highly science-driven. It provides new understanding into the 'demand-driven model of science', the role regulations play in determining the scientific research priorities. The study has examined this in the context of aftertreatment technologies that are specifically designed to reduce harmful emissions coming from a vehicle's tailpipe emissions. Aftertreatment systems are complex systems that needs integration of various technologies that are highly science driven. There is a high demand for the development of advanced aftertreatment systems as vehicle manufacturers have to strictly comply with emission regulations of the country where their vehicles are sold. The question that has been explored in the study is how emission regulations have shaped scientific research. Bibliometrics approach was applied to capture the temporal and longitudinal assessment of scientific activity over the period 1991-2023. The study thus also provides a novel approach to look into the relationship among two key research constructs. The vehicular regulation norms of Europe and USA have identified the pollutants that need to be controlled by vehicles with later evolution stages making it more stringent in terms of particle size and incorporation of additional pollutants that has to be controlled. The emission regulations in different countries have primarily been shaped by the emission standards of these two countries. The key research themes, topics/sub-topics of research over different periods (temporal analysis), and key areas research activity over the whole study period (longitudinal analysis) were identified. The study further drawing from this examined how the research activity mapped with the emission norms and more granularly with the pollutants that need to be controlled within a period. The paper has been able to demonstrate varied influences of emission regulations that have been implemented in phases over different periods in Europe and USA with the intensity and type of research activity.

The compliance with regulations calls for substantial changes in aftertreatment systems, such as the need for newer sophisticated filters, catalysts, and, in some

cases, radical changes in design configuration, among others. This pushes for investment in R&D to understand the scientific challenges with a focus on translational research. The research intensity overall and in different subcomponents, SCR, filters DPF, GPF, Lean NO_x and DOC catalysts reflects this. The introduction of newer pollutants and stringent limits imposed in later phases of the standards shows its influence on research across various key subcomponents of aftertreatment research. The key locations of research activity, the dispersion of research across different research areas, some key conferences playing a major role provides support to the argument of research activity influenced by regulations. The study identified the different types of catalysts, filters, sensors, reactants and reducing agents, specific technologies. Some of them were found to be more prominent and some newly introduced in later periods. Examination them with regulations standards demonstrated a closer correspondence with the emission regulations. It also showed contemporary areas where research is happening and where technology development is needed to address the implemented emission norms and newer norms like the Euro 7 (which is slated to be implemented soon) and Tier 4 US emissions that was implemented in 2024.

A study of this kind, however, does not capture the other aspects of research that are not reflected in research papers. In a science-driven technology field, it gives us a good indication of research influencing technology development. However, the innovation process does not simply start with pure scientific research and then progress linearly to technological development, but more of an intertwined relationship exists between science and technology throughout the innovation process (Branscomb, L. M., 2001). Patent citation studies provide a good indication of the highly mediated link between science and technology (Meyer, M. 2000). Thus, future studies on non-patent citations can help identify papers that have influenced technology development more directly than this study draws attention to. Future studies based on patents may bring some new insights into technology development influenced by regulations. However, in spite of this limitation, the study has shown how overall research has evolved over the years in aftertreatment systems for controlling tailpipe emissions. The overall research profile and at granular levels reveal to some extent the influence of regulation over different periods. The study's novelty lies in looking at regulation as a demand pull for scientific research and applying bibliometrics to capture this factor. The study inspite of its limitations has shown a tangible influence of emission regulations on scientific research.

References

Ai, W., Wang, J., Wen, J., Wang, S., Tan, W., Zhang, Z., Liang, K., Zhang, R. & Li, W. (2023). Research landscape and hotspots of selective catalytic reduction (SCR) for NOx removal: insights from a comprehensive bibliometric analysis. *Environmental Science and Pollution Research*, 30(24), 65482-65499.

Bhattacharya, S. & Basu, P. (1998). Mapping a research area at the micro level using co-word analysis. *Scientometrics*, 43(3), 359-372.

Branscomb. L.M. (2011). Technological Innovation. Editor(s): Neil J. Smelser, Paul B. Baltes. International Encyclopedia of the Social & Behavioral Sciences, Pergamon, USA.

Callon, M., Courtial, J. & Laville, F. (1991). Co-word analysis as a tool for describing the network of interactions between basic and technological research—The case of polymer chemistry. *Scientometrics*, 22(1), 155–205.

Callon, M., Courtial, J.P., Turner, W.A. & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2), 191–235.

Chen, X., Lun, Y., Yan, J., Hao, T. & Weng, H. (2019). Discovering thematic change and evolution of utilizing social media for healthcare research. *BMC Medical Informatics and Decision Making*, 19, 39-53.

Cobo MJ, Chiclana F, Collop A, de Oña J. & Herrera-Viedma E. (2014). A bibliometric analysis of the intelligent transportation systems research based on science mapping. *IEEE Trans Intell Transp Syst*; 15(2):901–8.

Cobo, M. J., Lopez-Herrera, A. G., Herrera-Viedma, E. & Herrera, F. (2011b). Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7), 1382–1402

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E. &Herrera, F. (2011a). An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics*, 5(1), 146–166.

Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E. & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, 63(8), 1609-1630.

Coulter, N., Monarch, I. & Konda, S. (1998). Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49(13), 1206-1223.

Crippa M., Janessensen-Moenhout, G., Guizzardi, D. & Galmarini, S. (2016). EU effect: Exporting emission standards for vehicles through the global market economy. *Journal of Environment Management*, 183, Part 3, 951-971.

Demers D. & Walters G. (1999). Guide to exhaust emission control options. BAeSAME, Bristol.

Egan, L., Mohammadpour, J. & Salehi, F. (2023). A bibliometric analysis of scientific research trends in monitoring systems for measuring ship emissions. *Environmental Science and Pollution Research*, 30(21), 60254-60267.

Ghosh, B., Dutta, S. P. &Mallik, A. (2020). Evolving Trends of Indian Research Performance in Cryptography: A Bibliometric and Computational Investigation. *Journal of Scientometric Research*, 9(3), 253–267.

Gajbhiye, M. D., Lakshmanan, S., Kumar, N., Bhattacharya, S. & Nishad, S. (2023). Effectiveness of India's Bharat Stage mitigation measures in reducing vehicular emissions. *Transportation Research Part D: Transport and Environmen*t, 115, 103603.

Ghosh, B., Dutta, S. P., & Mallik, A. (2020). Evolving Trends of Indian Research Performance in Cryptography: A Bibliometric and Computational Investigation. *Journal of Scientometric Research*, 9(3), 253–267.

Hamers, L., (1989). Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. *Information Processing and Management*, 25(3), pp.315-18.

Hiroyuki Y, Misawa K, Suzuki D, Tanaka K, Matsumoto J, Fujii M & Tanaka K (2011) Detailed analysis of diesel vehicle exhaust emissions: nitrogen oxides, hydrocarbons and particulate size distributions. Proc Combust Inst 33:2895–2902.

Huang MH, & Chang CP. (2014) Detecting research fronts in the OLED field using bibliographic coupling with sliding window. Scientometrics; 98(3): 1721–44.

Huang, L., Wang, X., Wu, F. & Li, Q. (2017). Analysis of Evolution and Frontier Research of Selective Catalyst Reduction Technology for Diesel Engine Based on Bibliometrics. In 2017 Portland International Conference on Management of Engineering and Technology (PICMET) (pp. 1-7). IEEE.

Frenkel, A., Maital, S., Leck, E. & Isreal. E. (2014). Demand-Driven Innovation: An Integrative Systems-Based Review of the Literature. International Journal of Technology and Management. 12-(2)

IIS (2014). International implementation of vehicle emissions standards https://www.climatechangeauthority.gov.au/reviews/light-vehicle-emissions-standards-australia/international-implementation-vehicle-emissions

JalaliFarahani V, Altuwayjiri A, Taghvaee S & Sioutas C. (2022). Tailpipe and non-tailpipe emission factors and source contributions of PM10 on major freeways in the Los Angeles basin. *Environmental Science & Technology*. Mar 1; 56 (11):7029-39.

Jones, W.P. & Furnas, G.W. (1987) Pictures of relevance: A geometric analysis of similarity measures. *Journal of the American Society for Information Science*, 38(6):420–442.

Joshi, A. (2020). Review of vehicle engine efficiency and emissions. *SAE International Journal of Advances and Current Practices in Mobility*, 2, 2479-2507.

Leydesdorff, L. & Persson, O. (2010). Mapping the geography of science: Distribution patterns and networks of relations among cities and institutes. *Journal of the American Society for Information Science and Technolo*gy, 61(8), 1622–1634.

Martinovic F, Andana T, Piumetti M, Armandi M. & Bonelli B (2020). Simultaneous improvement of ammonia mediated NOx SCR and soot oxidation for enhanced SCR-on-Filter application. *Applied Catalysis A: Genera*l. 596:117538.

Meyer, M. (2000). Does science push technology? Patents citing scientific literature. *Research Policy*, 29(3), 409-434.

Montero-Díaz, J., Cobo, M. J., Gutiérrez-Salcedo, M., Segado-Boj, F., and Herrera-Viedma, E. (2018). A science mapping analysis of 'Communication 'WoS subject category (1980-2013). *Comunicar: Revista Científica de Comunicacíon y Educacíon*, 26(55), 81-91.

Moral-Munoz, J.A., Arroyo-Morales, M., Herrera-Viedma, E. & Cobo, M.J. (2018). An Overview of Thematic Evolution of Physical Therapy Research Area From 1951 to 2013, *Front. Res. Metr. Anal.* 3 (March) 1–11.

Reşitoğlu, İ.A., Altinişik, K. & Keskin, A. (2015). The pollutant emissions from diesel-engine vehicles and exhaust aftertreatment systems. *Clean Techn Environ Po*licy 17, 15–27.

Rorvig, M. (1999). Images of similarity: A visual exploration of optimal similarity metrics and scaling properties of TREC topic-document sets. *Journal of the American Society for Information Science*, 50(8):639–651.

Ruegg, R. & Thomas, P. (2011). Linkages from DOE's Vehicle Technologies R&D in Advanced Combustion to More Efficient, Cleaner-Burning Engines (No. DOE/EE-0580). TIA Consulting Inc., Emerald Isle, NC (United States); 1790 Analytics, LLC, Haddonfield, NJ (United States).

Salton, G. & McGill, M.J. (1983). Introduction to modern information retrieval. McGraw-Hill.

Singh, S., Kulshrestha, M.J. & Rani, N. (2023). An Overview of Vehicular Emission Standards. *MAPAN* 38, 241–263.

Sternitzke C. & Bergmann I. (2009). Similarity measures for document mapping: a comparative study on the level of an individual scientist. *Scientometrics*; 78(1):113–30.

Tian, X., Geng, Y., Zhong, S., Wilson, J., Gao, C., Chen, W., Yu, Z. & Hao, H. (2018). A bibliometric analysis on trends and characters of carbon emissions from the transport sector. *Transportation Research Part D: Transport and Environment*, 59, 1-10.

Yu, R. C., Cole, A. S., Stroia, B. J., Huang, S. C., Howden, K. & Chalk, S. (2002). Development of Diesel Exhaust Aftertreatment System for Tier II Emissions. *SAE Transactions*, 111, 861–875.

World Health Organization. (2018). World Health Organization releases new global air pollution data. Retrieved from https://www.ccacoalition.org/en/news/world-health-organization-releases-new-global-air-pollution-dataLast accessed April 3, 2024.

https://doi.org/10.51408/issi2025_005

Insiders and Outsiders in International Scientific Collaboration: Distinguishing between Investigating and Investigated Countries

Zhe Cao[1], Lin Zhang[2], Gunnar Sivertsen[3], Zhihan Wan[4]

[1] *caozhe@whu.edu.cn*
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)

[2] *linzhang1117@whu.edu.cn*
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)
Centre for R&D Monitoring (ECOOM) and Department of MSI, KU Leuven, Leuven (Belgium)

[3] *gunnar.sivertsen@nifu.no*
Nordic Institute for Studies in Innovation, Research and Education (NIFU), Oslo (Norway)

[4] *zhihanwan@whu.edu.cn*
Center for Science, Technology & Education Assessment (CSTEA), Wuhan University, Wuhan (China)
School of Information Management, Wuhan University, Wuhan (China)

Abstract

This study focuses on patterns of international collaboration in cases where teams of scientists collaborate to conduct research aimed at solving problems pertinent to certain countries or regions. We employ Merton's insider-outsider theory to categorize authors from the countries under study as insiders and those from outside the studied countries as outsiders. We identify five collaboration patterns (CPs) based on different types of shared perspectives of co-authors – *Internal Perspective* (CP1), *Combined Perspective* (CP2), *Expanded perspective* (CP3), *Partially Overlapping Perspective* (CP4) and *External Perspective* (CP5). An empirical analysis of research related to "Sustainable Development Goal 1: No Poverty" reveals that CP1 is the most prevalent perspective. Whereas CP5 has seen a gradual decline, CP2 has risen over the years. A case study on the involvement of international scholars in poverty research in African countries reveals significant benefits from outsider participation, with substantial funding from developed countries. While this support has enhanced the quantity of research outputs, it also poses challenges. It may shape the perspectives and research agendas of insiders, thereby complicating internal efforts to develop research topics rooted in the local context and addressing domestic development needs.

Introduction

Research collaboration is a longstanding topic of interest in the field of science of science. Collaborators bring specialized knowledge and skills, each offering unique perspectives on research questions. By harnessing these strengths and fostering consensus among partners, collaboration often enhances efficiency and improves outcomes in scientific research. In the contemporary world, marked by pressing

challenges such as public health crises, climate change and energy sustainability, collaboration has become increasingly indispensable for tackling complex global problems. Investigating different ways to conduct scientific collaboration to figure out the effective collaboration patterns has thus emerged as a crucial topic of discussion among scholars.

Existing studies on research collaboration patterns predominantly emphasize the author aspect of collaborations. These studies typically categorize collaborations based on factors such as team size (e.g. large vs. small (Wu et al., 2019)), geographical scope (e.g. intra- vs. inter-institutional (Savić et al., 2017), domestic vs. international (Gök & Karaulova, 2024)), demographic attributes (e.g. gender (Love et al., 2022), ethnicity (AlShebli et al., 2018), professional status (Liu et al., 2019)), organizational structure (e.g. flat vs. hierarchical (Xu et al., 2022)), disciplinary backgrounds (e.g. disciplinary vs. interdisciplinary (Liu et al., 2024)) or the relational dynamics among collaborators (Feng & Kirkley, 2020). However, there remains a gap in addressing how these characteristics of authors correspond to the specific issues they aim to solve.

In response to this gap, the theory of insiders and outsiders (Merton, 1972) may offer a novel perspective for analyzing collaboration patterns. This theory posits that individuals can be classified as insiders or outsiders based on their alignment with societal norms, values and established rules within a specific context. Applied to research collaboration, it allows for the categorization of authors based on their alignment with the issues they study. This categorization may encompass various perspectives. For example, from a disciplinary perspective, authors can be classified as insiders or outsiders according to the degree of expertise in the field that the research problem belongs to. A typical research topic related to this perspective is interdisciplinarity, an area that has already been extensively explored in existing literature. However, this study adopts a geographical perspective by linking the origins of authors to the geographical focus of their research. This perspective corresponds to the growing emphasis on diverse contributions and practical solutions in scientific research evaluation (CoARA, 2022), which has led the research to increasingly address local issues to meet societal needs. Simultaneously, the complexity and integration of scientific problems make research collaboration a prevailing trend. In this context, how can different kinds of expertise and background contribute to solving specific problems that arise in local contexts? The insider-outsider theory provides valuable guidance for answering such questions. By exploring these dynamics, we move beyond traditional author-centric analyses to examine how diverse compositions of authors from different geographical backgrounds contribute to addressing geographically targeted problems.

On the background discussed above, this study addresses three main questions: (1) What collaboration patterns can be identified when viewed through the lens of insiders and outsiders? (2) Does the distribution of different collaboration patterns vary over time? (3) How do the topics of research vary across these collaboration patterns? We construct a new framework for identifying international scientific collaboration patterns, and utilize data from research related to the theme of

"Sustainable Development Goal 1: No Poverty" for the empirical analysis. The primary objectives are to elucidate evolutional trends and thematic features of outputs across various collaboration patterns. Furthermore, this study examines international academic activities aimed at poverty alleviation, with a particular focus on the engagement of Global North in the poverty research of Global South. It aims to offer insights to enhance collaborative research efforts and drive local solutions.

Theoretical framework

Theory of insiders and outsiders

In 1972, the American sociologist of science Robert Merton adopted a structural conception of insider/outsider status, defining insiders as "the members of specified groups and collectivities or occupants of specified social statuses" and outsiders as "the nonmembers" (Merton, 1972). The insider doctrine holds that "you have to be one in order to understand one". It posits that an individual has monopolistic or privileged access to knowledge, or is wholly excluded from it, by virtue of one's group membership or social position. According to this doctrine, the outsider has a structurally imposed incapacity to comprehend alien groups, statuses, cultures and societies. On the contrary, the outsider doctrine holds that "one need not to be Caesar in order to understand Caesar". It posits that individuals who are not bound by commitments to a specific group can readily assume the role of relatively objective investigators. In the fields of history and sociology, external perspectives can often provide profound insights and enhanced understanding. However, Merton holds the belief that achieving a transition from social conflict to intellectual controversy, wherein the perspectives of each group are taken seriously enough to be carefully examined rather than rejected out of hand, can facilitate a constructive interplay between the distinctive strengths and limitations of insider and outsider perspectives. This interplay, in turn, may enhance the potential for a more nuanced and comprehensive understanding of social life.

The theory of insiders and outsiders provides a proper perspective to revisit scientific collaboration in which authors with different affiliations and distinct characteristics work together to address specific scientific problems and co-publish their research findings. When the research problem pertains to a particular group, an author's status as an insider or outsider can be determined by his or her affiliation with that group. Insiders and outsiders may be contributing to the research target in different ways – insiders by possessing pre-existing membership within the group prior to the commencement of the research, and outsiders by entering the targeted context solely during the research process. Insiders and outsiders also may exhibit various research focuses. Insiders tend to prioritize the specific context and develop practical knowledge, whereas outsiders are more inclined to seek knowledge that can be generalized across various situations (Louis & Bartunek, 1992). Previous research has found that collaborative research has advantages for both insiders and outsiders, and for the nature of the research itself (Liu & Burnett, 2022). For outsiders, it allows easy access and achieves trust and acceptance by the local community. For insiders,

who may have some pre-formed biases that may influence their objectivity, they can be assisted by the outsider member of the team to retain a critical distance from the subject. This study adopts the structural conception of insider/outsider status and seeks to deepen the existing research by defining different collaboration patterns and delineating their distinctive characteristics.

Collaboration patterns from the insider-outsider perspective

In Merton's theory, the distinction between insider and outsider groups can be determined by various attributes such as gender, race, culture and region. This study provides an operational definition of insiders and outsiders in collaborative science from a geographical perspective. It categorizes authors from the countries under study as insiders and those from outside the studied countries as outsiders. Given that the typological classification is an effective means of understanding and interpreting phenomena (Bailey, 1994), this study categorizes different collaboration patterns (CPs) from the insider-outsider perspective. Specifically, we compare the ensemble of author countries (referred to as "investigating countries") and the ensemble of countries under study (referred to as "investigated countries") to define five CPs, as illustrated in Figure 1.

In this context, "author countries" refer to the nations of the institutions with which the authors are affiliated at the time of publishing the collaborative research. Conversely, "countries under study" pertain to the nations that are the focus of the research, such as the countries whose issues are being addressed or used as a research sample. For example, if scholars from the United Kingdom conduct research on economic development issues in South Africa, the investigating country would be the United Kingdom, while the investigated country would be South Africa.

The specific connotations of the five CPs are elucidated as follows.

- *CP1: Internal Perspective*
  Under this pattern, the investigating and investigated countries entirely coincide, indicating that researchers from specific countries focus on issues pertinent to their own nations. Such research typically embodies a distinct native perspective.

- *CP2: Combined Perspective*
  Under this pattern, the investigating countries encompass the investigated countries, indicating that domestic researchers engage in collaborative research with international counterparts to address domestic issues. Such research typically incorporates both internal and external perspectives to tackle local challenges.

- *CP3: Expanded perspective*
  Under this pattern, the investigated countries encompass the investigating countries, indicating that researchers from particular nations investigate issues relevant to both their own countries and other countries. Such research allows

for the examination of geographically extensive problems from specific perspectives.

- *CP4: Partially Overlapping Perspective*
  Under this pattern, the investigating and investigated countries exhibit intersections but do not completely overlap. The problems to be addressed and researchers' perspectives on problem-solving become more complicated.

- *CP5: External Perspective*
  Under this pattern, the investigating and investigated countries are entirely disjoint, indicating that researchers from particular nations investigate issues pertaining to other countries. Such research is often characterized by a completely external perspective.

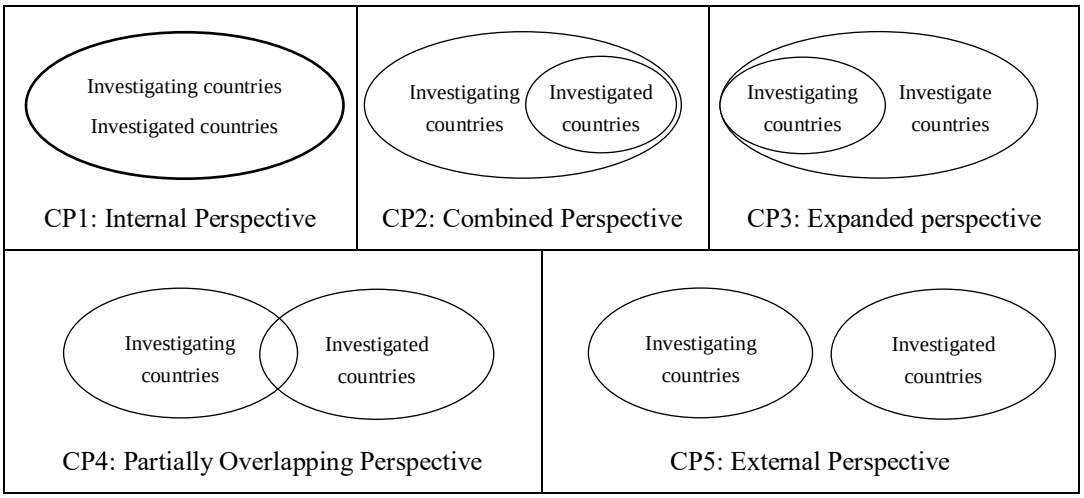


Figure 1. Five collaboration patterns from the insider-outsider perspective.

It should be noted that these five patterns do not encompass all types of scholarly papers. This study only analyzes papers that are identifiable to the author countries and focus on issues pertaining to certain countries. Actually, the framework for categorizing collaboration patterns proposed in this study is topic-dependent and therefore particularly well-suited for research addressing issues within health, environment, humanities and social sciences, where the emphasis is more on studying problems in geographical contexts. In contrast, its applicability is relatively constrained in many physical science fields that prioritize the identification of universal scientific laws.

Data and method

Data

This study takes academic papers related to Sustainable Development Goal 1 as cases to conduct empirical analysis. The Sustainable Development Goals (SDGs) were adopted by the United Nations in 2015 as a universal call to action to end poverty, protect the planet and ensure that by 2030 all people enjoy peace and prosperity. Among 17 goals in this 15-year plan, the first goal "SDG1 No Poverty" aims to end poverty in all its forms everywhere. Scholarly investigations pertaining to SDG1 are more likely to focus on country-specific contexts, thereby closely aligning with the requisites of this study.

At the operational level, Elsevier has generated SDG search queries to help researchers and institutions track and demonstrate progress toward the SDG targets since 2018 (Scopus, 2023). These queries, along with the university's own data and evidence supporting progress and contributions to the particular SDG outside of research-based metrics, have been used for the THE Impact Rankings. The latest 2023 SDG queries are a result of Elsevier data science teams building extensive keyword queries, supplemented with a predictive machine learning element, to map documents to SDGs with very high precision (Bedard-Vallee et al., 2023). Employing the newest version of queries provided by Elsevier, this study downloaded 223,816 papers (including the document types of Article and Review) related to SDG1 from Scopus (https://www.scopus.com/). The data was retrieved in May 2024.

In addition to the bibliographic data obtained from the Scopus database, this study also incorporates extensive metrics from the SciVal platform (https://www.scival.com/), an analytical tool developed by Elsevier based on Scopus data. Detailed descriptions of the application of these metrics will be presented in the subsequent sections where they are used in our analysis.

Method

From a technical perspective, the challenging aspect of this study lies in the identification of the investigating countries and the investigated countries. For the former, the country entities are extracted from the structured list of author affiliations provided in the bibliographic information of papers using regular expressions. For the latter, the country entities are extracted from the titles, author keywords and abstracts provided in the bibliographic information of papers using the *spaCy*, a free open-source library for natural language processing in Python. Subsequently, the country names are standardized using the *pycountry* library. Once the investigating countries and the investigated countries are determined, the collaboration pattern of each paper can be identified.

Here, two issues require clarification. Firstly, regarding the identification of the investigating countries, the institutions to which the authors are affiliated may not always accurately reflect their native cultural groups. For instance, some authors studying or visiting abroad may be affiliated with institutions from both their home

and host countries. Nonetheless, considering that these transnational authors possess a certain degree of cultural perspective from the host country, identifying the investigating countries through the authors' affiliations is still deemed reasonable.
Secondly, concerning the identification of the investigated countries, the mention of a country in the title, keywords or abstract – especially when only mentioned in the abstract – does not necessarily imply that the research focuses on issues specific to that country or is based on its real-world conditions. It may merely use the country as a research context. Moreover, sometimes only a city or region within a country or a country group is mentioned without referencing the country itself. However, after manually checking 200 pieces of abstracts, it was found that less than 5% of cases resulted in erroneous collaboration pattern identification due to the aforementioned reasons. Thereby, the method used in this study for determining the collaboration pattern is considered to be fairly precise.
In the overall sample, 209,570 papers (93.6%) contain information of author affiliation and include at least one field among the title, author keywords and the abstract. Since not all studies center around specific research subjects, 112,110 papers (53.5%) with identifiable collaboration patterns from the insider-outsider perspective are selected for the following analysis in this study.

Results

Panoramic view: Distribution and features of five collaboration patterns

This study commences with an extensive data analysis of the sample to reveal the collaborative characteristics of research on poverty issues. The key findings in this section are as follows: *Internal Perspective* is the most prevalent collaboration pattern overall; research under the pattern of *External Perspective* has gradually decreased over time, while that of *Combined Perspective* has increased. A general finding is also that research incorporating an outsider perspective focuses on more cutting-edge topics.

Overview

Among five collaboration patterns, *Internal Perspective* (CP1) is the most commonly-observed one, with 57,687 (51.5%) pieces of papers in total. Patterns of *External Perspective* (CP5) and *Combined Perspective* (CP2) are also prevalent, with 25,936 (23.1%) and 20,371 (18.2%) pieces of papers respectively. Patterns of *Expanded perspective* (CP3) and *Partially Overlapping Perspective* (CP4) are relatively rare, with 6,236 (5.6%) and 1,880 (1.7%) pieces of papers respectively.
Considering the specific distribution across countries, certain differences can be observed across different collaboration patterns in terms of the investigating countries and the investigated countries (see Table 1). The primary finding is that, in addressing the issue of poverty, developed countries are more inclined to act as

initiators of research, while developing countries are more frequently the focuses of these studies[1].

Table 1. Numbers and proportions of representative country combinations in five collaboration patterns[2].

CP1: Internal Perspective			*CP2: Combined Perspective*				*CP3: Expanded perspective*			
A/B	**N**	**P**	**B**	**A-B**	**N**	**P**	**A**	**B-A**	**N**	**P**
US	7175	12.4%	CN (2532, 12.4%)	US	855	33.8%	US (996, 16.0%)	IN	186	18.7%
CN	6052	10.5%		GB	294	11.6%		MX	131	13.2%
IN	4043	7.0%		AU	274	10.8%		CA	65	6.5%
GB	3877	6.7%		CA	115	4.5%		others	614	61.6%
ZA	3084	5.3%		others	994	39.3%	GB (536, 8.6%)	US	113	21.1%
AU	2173	3.8%	IN (1184, 5.8%)	US	292	24.7%		DE	25	4.7%
BR	2108	3.7%		GB	163	13.8%		others	398	74.3%
CA	1849	3.2%		AU	86	7.3%	CN (270, 4.3%)	US	33	12.2%
DE	1624	2.8%		NL	38	3.2%		JP	25	9.3%
ES	1488	2.6%		others	605	51.1%		SG	16	5.9%
ID	1218	2.1%	ZA (933, 4.6%)	GB	214	22.9%		others	196	72.6%
NG	1192	2.1%		US	177	19.0%	ZA (253, 4.1%)	ZW	23	9.1%
IT	1052	1.8%		NL	37	4.0%		IN	18	7.1%
SE	985	1.7%		NG	35	3.8%		US	12	4.7%
MX	909	1.6%		others	470	50.4%		others	200	79.1%
others	18858	32.7%	others (15722, 77.2%)				others (4181, 67.0%)			

CP4: Partially Overlapping Perspective									*CP5: External Perspective*			
A&B	**N**	**P**	**A-B**	**N**	**P**	**B-A**	**N**	**P**	**A**	**B**	**N**	**P**
CN	109	5.8%	US	307	16.3%	US	104	5.5%	US	IN	591	2.3%
US	86	4.6%	GB	209	11.1%	IN	77	4.1%	US	CN	557	2.1%
ZA	80	4.3%	CA	73	3.9%	CN	64	3.4%	US	MX	423	1.6%
GB	77	4.1%	AU	70	3.7%	GB	51	2.7%	GB	IN	311	1.2%
IN	68	3.6%	NL	66	3.5%	DE	37	2.0%	US	ES	258	1.0%
DE	53	2.8%	DE	53	2.8%	ES	33	1.8%	GB	CN	219	0.8%
BR	46	2.4%	FR	36	1.9%	CA	24	1.3%	US	BR	202	0.8%
KE	41	2.2%	GB, US	34	1.8%	FR	23	1.2%	US	ZA	191	0.7%
PK	41	2.2%	ZA	29	1.5%	AU	21	1.1%	GB	US	157	0.6%
ES	36	1.9%	IT	28	1.5%	JP	21	1.1%	US	GE	156	0.6%
BD	34	1.8%	ES	27	1.4%	KE	21	1.1%	GB	ZA	153	0.6%
IT	33	1.8%	CH	23	1.2%	SE	21	1.1%	ZA	ZW	151	0.6%
AU	32	1.7%	CN	23	1.2%	PT	20	1.1%	US	BD	146	0.6%
GH	32	1.7%	SE	23	1.2%	BR	18	1.0%	AU	CN	143	0.6%
MX	32	1.7%	BE	21	1.1%	GH	17	0.9%	US	KE	134	0.5%
others	1080	57.4%	others	858	45.6%	others	1328	70.6%	others		22144	85.4%

*** A: investigating country ensemble; B: investigated country ensemble; N: number; P: proportion; /: or; &: and; -: except.**

[1] As of 2023, there are 37 globally recognized developed countries acknowledged by institutions such as the World Bank, the International Monetary Fund, the United Nations Development Programme, and the Central Intelligence Agency of the United States. These countries include the United Kingdom, France, Germany, Italy, the Netherlands, Norway, Sweden, Finland, Denmark, Iceland, Switzerland, Belgium, Luxembourg, Ireland, Spain, Portugal, Austria, the Czech Republic, Slovakia, Hungary, Greece, Slovenia, Poland, Estonia, Latvia, Lithuania, Malta, San Marino, Cyprus, Japan, South Korea, Singapore, Israel, the United States, Canada, Australia, and New Zealand.

[2] To save the space, the binary codes of countries are employed in this study. The binary codes and the corresponding full names are detailed in Table 5 in the appendix.

Regarding each specific pattern, 98.4% CP1 papers only cover one investigating and investigated country, which means that the authors are mostly from a single country and they study issues related to their own country. The countries with the highest numbers of CP1 papers are the United States (7,175, 12.4%), China (6,052, 10.5%), India (4,043, 7.0%), the United Kingdom (3,877, 6.7%) and South Africa (3,084, 5.3%). CP2 papers feature the collaboration between developing and developed countries in studying issues pertinent to developing countries. Among these papers, 97.4% only have one investigated country, with three most focused countries being China (2,532, 12.4%), India (1,184, 5.8%) and South Africa (933, 4.6%). CP3 papers feature diverse investigating and investigated countries. Among these papers, 95.6% only have one investigating country, with two most active countries being the United States (996, 16.0%) and the United Kingdom (536, 8.6%). China (270, 4.3%) and South Africa (253, 4.1%) are also important investigating countries. The countries under study are diverse, exhibiting various characteristics such as cultural similarity, geographical proximity and comparable levels of development. CP4 papers involve three types of countries – intersections of the ensembles of investigating countries and investigated countries, countries only in the investigating country ensembles, and countries only in the investigated country ensembles. Notably, the developing countries appear more in the intersections. Among CP5 papers, the pairs of investigating countries and investigated countries, which indicate who study whom, are worth the attention. Although the distribution of country pairs is relatively dispersed, a clear pattern emerges: authors mostly come from developed countries, while the research primarily focuses on issues pertaining to developing countries.

Temporal trend

By further examining the trends over the years (see Figure 2), it is evident that the proportion of CP1 papers has remained stable over the past two decades. The most significant change is the shift from studying issues in other countries from an outsider's perspective to engaging in collaborative research between insiders and outsiders. The proportion of CP2 papers has increased from 7.6% in 2000 to 22.2% in 2023, while the proportion of CP5 papers has decreased from 31.4% in 2000 to 16.5% in 2023. When combined with country-level information, this trend suggests a growing collaboration between the global North and South in addressing poverty-related issues.

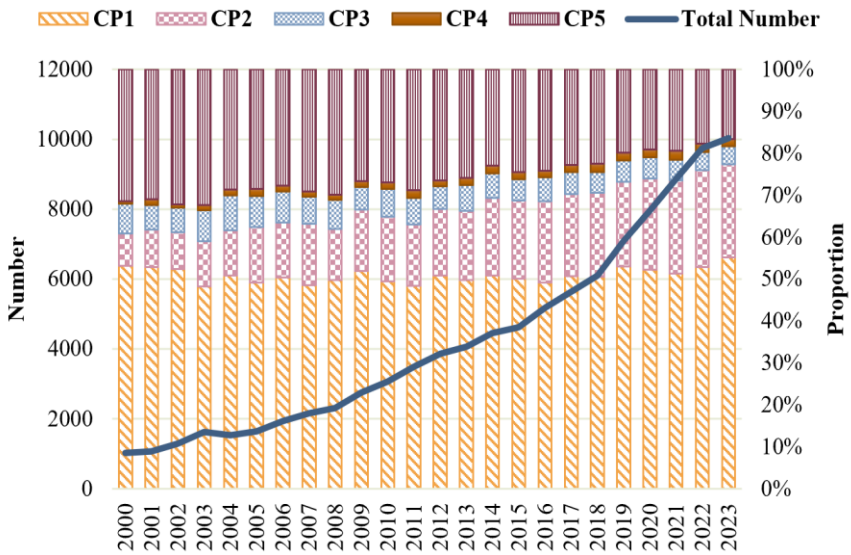


Figure 2. Annual trend of the total number of papers and the proportion of papers under five collaboration patterns.

Note: Due to the relatively low number of papers in certain years, to ensure the clarity and aesthetic quality of the figure, only papers published within 2000~2023 are displayed, covering over 90% of the overall samples.

Thematic feature

In terms of thematic features (see Figure 3), this study examines two indicators based on the topics annotated for individual papers by the SciVal platform – the *Topic Prominence Percentile*, a metric provided by SciVal reflecting the momentum of the topic; and the *Topic Diversity*, a self-developed indicator that calculates the diversity of topics using the Simpson index (Simpson, 1949).

Regarding the topic prominence, collaboration patterns integrating external perspectives or involving partial engagement exhibit relatively high average levels of topic prominence, while collaboration patterns that mainly rely on internal perspectives show relatively lower average levels of topic prominence. Regarding the thematic diversity, the pattern with insiders self-looking demonstrates the highest indicator level all the time. In contrast, other collaboration patterns incorporating external perspectives initially exhibit relatively low topic diversity, which increases over time.

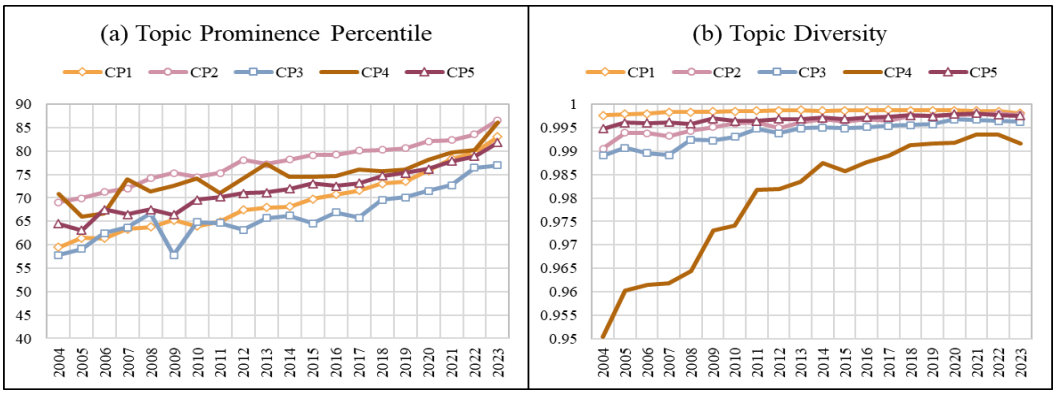


Figure 3. Indicator level of thematic features for five collaboration patterns.

Case study: International engagement in poverty research in Africa

Africa represents the youngest per person and fastest-growing population in the world, with the oldest and most diverse genome (Marincola & Kariuki, 2020). However, poverty has long been a central issue in African development due to factors such as inadequate economic growth, poor governance, cultural challenges, conflict and disease (Omomowo, 2018). Until now, Africa remains "the core of the world's poverty problem" (Bigman, 2011). Historically, the interaction among African countries is relatively limited, which is particularly pronounced when compared to partnerships with more developed regions such as Europe, Asia and America (Dine et al., 2024). Instead, research in African countries has largely been conducted by scholars from the Global North (Vieira, 2022). However, to study Africa effectively, it is essential to develop a comprehensive understanding of the region (Dine et al., 2024). As African countries are experiencing a shift towards more equitable and sustained research partnerships (Eduan & Yuanqun, 2019; Vieira, 2022), it is crucial to examine the contributions of both internal and external actors in the poverty research in Africa. This section zooms into 19,437 research articles with poverty in African countries as the topic.

Who are the insiders? Who are the outsiders?

At the outset of this case study, it is essential to clarify again the definitions of "insiders" and "outsiders". In the prior analysis, different collaboration patterns were distinguished with nations as the basis for the units. However, our framework can be applied to any geographical unit, and a regional perspective covering groups of countries is adopted in this section. According to the five regions in Africa[3] – Eastern Africa, Southern Africa, Western Africa, Northern Africa and Central Africa, authors from within a specific region will be considered as insiders, while those from outside the region are regarded as outsiders. This is based on the assumption that people from the same African region may share relatively similar cultural backgrounds and research environments.

[3] The regional division of African countries is detailed in Table 6 of the Appendix.

When considering the five regions as investigated units, Eastern Africa is the most investigated (7,581, 39%), followed by Southern Africa (5,615, 28.89%) and Western Africa (5,608, 28.85%). Northern Africa (994, 5.11%) and Middle Africa (600, 3.09%) have been investigated relatively less from the same perspective of poverty. In Eastern Africa, countries including Ethiopia (1,802, 23.77%), Kenya (1,642, 21.66%), Tanzania (1,235, 16.29%), Uganda (1,188, 15.67%) and Zimbabwe (809, 10.67%) have received considerable attention; in Western Africa, relevant research is mostly concentrated on Nigeria (2,405, 42.83%) and Ghana (2,160, 38.47%); in Southern Africa, South Africa (5,058, 90.19%) stands out prominently; in Northern Africa and Central Africa, Egypt (400, 40.24%) and Cameroon (355, 59.17%) are respectively the most investigated country in their areas.

For all five regions, the United States and the United Kingdom are the main outsiders investigating into their poverty issues. As shown in Table 2, these two countries have participated in the highest share of papers outside of the region itself. Countries such as Canada, Germany, the Netherlands and Australia are also among those that have conducted extensive research on poverty in Africa. In particular, France demonstrates a relatively high level of attention towards issues pertaining to Northern and Central Africa.

Table 2. Top 10 investigating countries (regions) for five African regions and the share of their papers.

Overall (19,437)		Eastern Africa (7,581)		Western Africa (5,615)	
Investigator	Share	Investigator	Share	Investigator	Share
Southern Africa	26.02%	Eastern Africa	57.62%	Western Africa	65.95%
Eastern Africa	12.37%	US	23.86%	US	16.46%
Western Africa	11.11%	GB	18.61%	GB	13.96%
US	9.27%	Southern Africa	11.23%	Southern Africa	6.95%
GB	8.45%	DE	5.37%	CA	6.27%
CA	6.35%	NL	4.99%	Eastern Africa	4.93%
DE	6.11%	CA	4.21%	AU	4.26%
NL	4.16%	NO	3.39%	DE	3.60%
Northern Africa	2.34%	SE	3.22%	NL	3.26%
AU	2.06%	Western Africa	2.95%	FR	3.08%
Southern Africa (5,608)		**Northern Africa (994)**		**Central Africa (600)**	
Investigator	Share	Investigator	Share	Investigator	Share
Southern Africa	82.28%	Northern Africa	57.04%	Central Africa	47.17%
GB	11.82%	US	15.90%	US	16.67%
US	11.00%	GB	11.27%	GB	13.50%
Eastern Africa	2.92%	FR	9.56%	Southern Africa	10.33%
CA	2.91%	CA	3.72%	Eastern Africa	10.00%
Western Africa	2.64%	IT	3.52%	FR	9.17%
NL	2.41%	DE	3.42%	Western Africa	8.83%
AU	2.19%	Western Africa	3.12%	CA	7.33%
DE	2.07%	Eastern Africa	2.82%	DE	7.17%
SE	1.59%	SA	2.82%	BE	6.00%

To what extent do outsiders engage in insiders?

Figure 4 illustrates the temporal changes in the number of papers focused on countries in different African regions over the past two decades, as well as the proportion of papers under different collaboration patterns. The main observation of this subsection is that the poverty research in Africa is highly dependent on outsiders, with a growing trend toward collaborative research between insiders and outsiders.

For all papers addressing African poverty issues, it is consistent with the overall trend shown in Figure 2 that the proportion of papers under CP5 has decreased, while those under CP2 have risen. Different from the results shown in the overall sample, papers under CP1 are relatively scarce in African poverty research, particularly in the earlier years, with studies involving outsiders accounting for over 60% of the total.

Focusing on different African regions, the dependence on external scientific research forces is particularly prominent in Eastern and Central Africa, whereas Southern Africa exhibits stronger autonomy in conducting related research, with the proportion of CP1 papers exceeding 60%. Notably, in contrast to the prominent trend of other regions engaging in collaborative or independent research, the proportion of CP1 papers in Central Africa exhibits a declining trend, with an increasing reliance on outsider contributions instead.

Figure 4. Annual trend of the number of papers investigating different African regions and the distribution of five collaboration patterns.

How do the outsiders shape the research topics?

Given the substantial involvement of outsiders in the study of poverty in African regions, we will now analyze whether the research topics vary depending on whether the study is conducted by insiders alone (CP1), outsiders alone (CP5), or through collaboration between insiders and outsiders (CP2). Table 3 presents a comparison of the most frequently occurring topic keywords across three patterns for five African regions. These topic keywords are derived from the topic cluster names provided by the SciVal platform for each paper.

Generally, themes related to finance, industry, health and climate are the most investigated. Research conducted independently by insiders and outsiders demonstrates relatively consistent topic preferences, with a tendency to focus on economic and climate-related issues. In contrast, research jointly conducted by insiders and outsiders shows a clear focus on topics in the field of healthcare and medicine. Among these themes, finance is intrinsically and obviously linked to poverty; industrial development can alleviate poverty by promoting economic growth; the existence of health problems can be attributed to the pernicious cycle between disease and poverty; and environmental issues exacerbate poverty, because the impacts of climate change on food insecurity, forced migration, disease and mortality may bring African countries that are already vulnerable with increasingly severe and inequitable disasters.

A notable distinction is that CP1 and CP5 papers focused on Southern African countries tends to emphasize political and historical topics, such as democracy and colonialism. Meanwhile, CP5 papers focused on Northern African countries shows greater attention to religious and cultural issues, such as Islam and Arab culture, although the proportion of these papers is declining. Moreover, CP2 papers focused on Central African countries predominantly addresses environmental protection topics, such as natural resources, deforestation and environmental policies. These locally distinctive issues merit attention, which may offer unique insights for the international community.

Table 3. Proportion of papers with high-frequency topic keywords under different collaboration patterns.

Investigated region	CP1		CP2		CP5	
	Topic keyword	Share	Topic keyword	Share	Topic keyword	Share
Overall	Finance	9.5%	Health Service	14.5%	Finance	11.4%
	Climate Change	9.1%	Climate Change	11.8%	Climate Change	9.6%
	Income Inequality / Wealth	7.5%	Neonatal Infant	8.8%	Democracy	8.5%
			Finance	6.6%	Income Inequality / Wealth	8.5%
	Industry	7.2%	Mental Health	5.7%		
Eastern Africa	Climate Change	12.9%	Health Service	23.1%	Finance	11.9%
	Health Service	9.6%	Climate Change	19.1%	Climate Change	10.9%
	Neonatal Infant	7.9%	Neonatal Infant	15.5%	Income Inequality	8.6%

	Finance	7.7%	Natural Resource	10.5%	Wealth	8.6%
	Industry	6.7%	Toddlers	9.1%	Democracy	6.6%
Western Africa	Finance	10.2%	Health Service	18.9%	Finance	10.9%
	Health Service	9.6%	Climate Change	11.9%	Health Service	9.2%
	Climate Change	9.4%	Neonatal Infant	9.8%	Climate Change	9.0%
	Industry	9.2%	Delivery of Health Care	9.7%	Income Inequality / Industry / Wealth	7.9%
	Income Inequality / Wealth	7.3%	Household	9.3%		
Southern Africa	Democracy	9.3%	Health Service	10.3%	Democracy	16.1%
	Finance	9.2%	Finance	8.8%	Colonialism	11.8%
	Income Inequality / Wealth	8.3%	Income Inequality / Wealth	8.3%	Finance	11.5%
	Welfare	7.5%	Climate Change	8.2%	Income Inequality / Wealth	10.5%
Northern Africa	Finance	16.3%	Climate Change	19.4%	Islam	15.0%
	Income Inequality / Wealth	12.5%	Finance	8.8%	Democracy	14.4%
	Industry	10.3%	Income Inequality / Irrigation / Water Management / Wealth	7.4%	Finance	11.1%
	Democracy / Health Service / Social Media	7.2%			Industry	10.3%
					Arab World / Climate Change	9.5%
Central Africa	Finance	16.0%	Natural Resource	15.5%	Finance	10.6%
	Industry	11.0%	Climate Change	11.3%	Democracy / Industry	9.9%
	Income Inequality / Wealth	10.0%	Deforestation	10.7%		
			Health Service	10.1%	Climate Change / Natural Resource	8.0%
	Climate Change	8.0%	Environmental Policy	9.5%		

Collaboration patterns and sources of funding

In actual, the advancement of scientific research relies heavily on science funding, especially for research fields with substantial expenditures on instruments, materials, etc. To a certain extent, the choice of research topics is significantly influenced by the funding agencies. In particular, research funding plays an important role in shaping scientific collaborations between the North and the South (Skupien & Rüffin, 2019). Therefore, the second part of analysis in this subsection examines the participation of outsiders in poverty research in African regions from the perspective of science funding.

According to data provided by Scopus, among the 19,437 research articles with African countries as investigated countries, 6,852 (35%) of them are labeled with funding information. This proportion aligns with the overall sample, as only 38,096 out of 112,110 papers (34%) have funding information. Table 4 showcases the funding agencies with the highest number of associated publications in the overall case sample and papers investigating different African regions.

Generally, indigenous funding institutions in African countries are relatively limited, whose effects are only manifested in studies that exclusively include insiders. In contrast, grants from foundations in the United Kingdom, the United States, and other countries have played a significant role in advancing research on poverty in Africa. On the regional side, the National Research Foundation (NRF) of South Africa is the primary source of funding for research in Africa, while the Economic Research Forum (ERF) in Egypt, the African Development Bank, and universities in several African countries have also played a significant role in the production of CP1 papers. On the international side, international funding sources generally fall into three categories – institutions focused on international development, e.g., the UK Department for International Development (DFID), the US Agency for International Development (USAID) and the Canadian International Development Research Centre (IDRC); those concentrating on economic and social issues, e.g., the UK Economic and Social Research Council (ESRC) and the World Bank Group (WBG); and those specializing in medical research, e.g., the US National Institutes of Health (NIH). Such distribution of funding sources aligns with the thematic focus on finance, climate and health-related issues to a certain extent.

It should be noted that the absolute values presented in Table 4 reflect the primary institutions funding research on poverty in Africa but fail to adequately capture the level of attention these institutions devote to the issue of poverty in Africa. We have conducted a search in the Scopus database for the major funding agencies supporting global research under "SDG1 No Poverty". It has been found that, while institutions such as the ESRC, NSF, and NIH fund a considerable proportion of research on poverty in Africa, their contributions account for only 14.6%, 4.9% and 4.8%, respectively, of their total funding for global poverty research. In contrast, agencies like DFID, USAID and IDRC have 56.9%, 42.4% and 42.2% of their poverty research focused on African countries, respectively, demonstrating a distinctive focus on Africa by these agencies.

Table 4. Proportion of papers with high-frequency funding agencies under different collaboration patterns.

Investigated region	CP1		CP2		CP5	
	Funding agency	Share	Funding agency	Share	Funding agency	Share
Overall	NRF, ZA*	13.6%	DFID, UK	7.6%	ESRC, UK	7.3%
	DFID, UK	3.2%	EC	6.7%	DFID, UK	6.3%
	IDRC, CA	3.2%	USAID, US	6.0%	USAID, US	5.4%
	USAID, US	3.1%	NIH, US	5.6%	EC	4.9%
	Sida, SE	2.7%	BMGF, US	5.6%	WBG	4.8%
Eastern Africa	USAID, US	6.2%	DFID, UK	8.0%	ESRC, UK	8.6%
	Sida, SE	6.0%	BMGF, US	7.3%	DFID, UK	8.5%
	AAU, ET*	5.8%	USAID, US	7.0%	USAID, US	6.2%
	CREA*	4.3%	NIH, US	6.1%	WBG	5.1%

	DFID, UK	4.1%	EC	6.0%	EC	4.4%
Western Africa	CoU, NG*	6.9%	USAID, US	7.2%	USAID, US	6.2%
	IDRC, CA	6.9%	DFID, UK	7.1%	WBG	5.6%
	WBG	5.5%	BMGF, US	7.0%	ESRC, UK	4.9%
	DFID, UK	5.2%	EC	4.5%	DFID, UK	4.1%
	USAID, US / UCC, GH*	2.9%	IDRC, CA	4.1%	IDRC, CA	4.0%
Southern Africa	NRF, ZA*	29.6%	NRF, ZA	16.8%	ESRC, UK	10.5%
	WRC, ZA*	5.1%	ESRC, UK	11.7%	EC	6.6%
	UCT, ZA*	4.7%	EC	10.7%	DFID, UK	5.6%
	SAMRC, ZA*	3.8%	NIH, US	9.4%	SSHRC, CA	4.3%
	UJ, ZA*	3.0%	WT, UK	8.7%	NSF, US	3.9%
Northern Africa	ERF, EG*	11.5%	EC	13.5%	EC	8.4%
	IDB	7.7%	ERF, EG*	4.5%	ESRC, UK	4.7%
	CaU, EG*	5.8%	DFID, UK	3.4%	USAID, US	4.7%
	UNICEF	5.8%	MHESR, EG*	3.4%	ANR, FR	3.7%
Central Africa	IDRC, CA	15.0%	EC	12.3%	EC	11.7%
	ADBG* / CIFOR / WBG / WRI	10.0%	DFID, UK	11.0%	USAID, US	5.3%
			IDRC, CA	8.2%	ESRC, UK	4.3%
			USAID, US	6.9%	WBG	4.3%

Note: (1) * indicates African institutions. (2) The full names of the funding institutions can be found in Table 7 of the appendix.

Conclusion and discussion

This study introduces the sociological theory of insiders and outsiders into the context of scientific collaboration, and proposes five distinct collaboration patterns based on different types of shared perspectives of co-authors – *Internal Perspective* (CP1), *Combined Perspective* (CP2), *Expanded perspective* (CP3), *Partially Overlapping Perspective* (CP4) and *External Perspective* (CP5). It adopts academic papers related to "Sustainable Development Goal 1: No poverty", a topic characterized by significant contextual features, to conduct empirical analysis. The findings reveal that, the *Internal Perspective* has been the predominant collaboration pattern. However, in recent years, there has been a noticeable increase in research under the pattern of *Combined Perspective*. Research incorporating the outsider perspective tends to address more emerging topics. Collaborating on poverty research in specific countries or regions is becoming a prevailing trend. This approach serves as a crucial means for insiders to enhance their research capabilities, while it also offers outsiders an opportunity to gain in-depth contextual understanding and make substantial contributions. Theoretically, this study deepens and extends the research perspectives on scientific collaboration by looking more deeply into how different constellations are related to different topics.

More importantly, our case study focuses on the involvement of international scholars in poverty research within African countries, thus endowing the research

with significant practical relevance. The findings reveal that, with the exception of Southern Africa, with the National Research Foundation of South Africa serving as an essential funding source, the majority of poverty research in African regions largely depends on international contributions of competences and resources. While the engagement of outsiders can significantly expand the topics of the research, it is important to recognize that the lack of local leadership may dilute the local relevance of the research topics, shifting them towards more internationalized issues. This situation is partly attributable to insufficient domestic funding for scientific research, particularly from government sources. In contrast, countries such as the United Kingdom, the United States, and Canada have established dedicated government funding agencies targeting on international development and private institutions in specialized fields like medicine, which have played a crucial role in supporting research and solutions for poverty in the Global South. This reflects the positive contributions of external researchers, but it also highlights the need for local researchers to be aware of the potential loss of local discourse authority due to over-reliance on external support. It might be crucial for African countries to increase investment in scientific research and achieving technological self-reliance.

It should be recognized that while our sources of data can provide insights into the outcomes of collaborations between insiders and outsiders, they offering only limited understanding of the motivations behind the research collaborations. Given our focus on developing a new framework for categorizing and analyzing collaboration patterns from the insider-outsider perspective, deeper issues will warrant further examination. For instance, how do the research perspectives of insiders and outsiders mutually shape one another? What are the underlying mechanisms through which scientific funding impacts research topics? What are the similarities and differences in the academic impact and societal value of research outcomes produced by different collaboration patterns? These questions represent important areas for future investigation.

Acknowledgments

The authors would like to acknowledge support from the National Natural Science Foundation of China (Grant Nos. 72374160, L2424104) and the National Laboratory Centre for Library and Information Science at Wuhan University.

References

AlShebli, B. K., Rahwan, T., & Woon, W. L. (2018). The preeminence of ethnic diversity in scientific collaboration. *Nature Communications*, 9(1), 5163.

Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques*. Thousand Oaks, CA, US: Sage Publications, Inc.

Bedard-Vallee, A., James, C., & Roberge, G. (2023). Elsevier 2023 sustainable development goals (SDGs) mapping. Elsevier Data Repository, V1, doi: 10.17632/y2zyy9vwzy.1.

Bigman, D. (2011). *Poverty, hunger, and democracy in Africa: Potential and limitations of democracy in cementing multiethnic societies*. Basingstoke: Palgrave Macmillan London.

Dine, R. D., Elkheir, L. Y. M., Raimi, M. O., et al. (2024). Ten simple rules for successful and sustainable african research collaborations. *PLoS Computational Biology*, 20(6), e1012197.

Eduan, W., & Yuanqun, J. (2019). Patterns of the China-Africa research collaborations from 2006 to 2016: A bibliometric analysis. *Higher Education*, 77(6), 979-994.

Feng, S., & Kirkley, A. (2020). Mixing patterns in interdisciplinary co-authorship networks at multiple scales. *Scientific Reports*, 10(1), 7731.

Gök, A., & Karaulova, M. (2024). How "international" is international research collaboration? *Journal of the Association for Information Science and Technology*, 75(2), 97-114.

Liu, J., Ding, K., Wang, F., et al. (2019). The structure and evolution of scientific collaboration from the perspective of symbiosis. *Malaysian Journal of Library & Information Science*, 24, 59-73.

Liu, X., Bu, Y., Li, M., et al. (2024). Monodisciplinary collaboration disrupts science more than multidisciplinary collaboration. *Journal of the Association for Information Science and Technology*, 75(1), 59-78.

Liu, X., & Burnett, D. (2022). Insider-outsider: Methodological reflections on collaborative intercultural research. *Humanities and Social Sciences Communications*, 9(1), 314.

Louis, M. R., & Bartunek, J. M. (1992). Insider/outsider research teams: Collaboration across diverse perspectives. *Journal of Management Inquiry*, 1(2), 101-110.

Love, H. B., Stephens, A., Fosdick, B. K., et al. (2022). The impact of gender diversity on scientific research teams: A need to broaden and accelerate future research. *Humanities and Social Sciences Communications*, 9(1), 386.

Marincola, E., & Kariuki, T. (2020). Quality research in Africa and why it is important. *ACS Omega*, 5(38), 24155-24157.

Merton, R. K. (1972). Insiders and outsiders: A chapter in the sociology of knowledge. *American journal of Sociology*, 78(1), 9-47.

Omomowo, K. E. (2018). Poverty in Africa. In O. Akanle & J. O. Adésìnà (Eds.), *The development of Africa: Issues, diagnoses and prognoses* (pp. 69-94). Springer International Publishing.

Savić, M., Ivanović, M., & Dimić Surla, B. (2017). Analysis of intra-institutional research collaboration: A case of a Serbian faculty of sciences. *Scientometrics*, 110(1), 195-216.

Scopus. (2023, 2023-08-21). What are sustainable development goals (SDGs)? Retrieved 2024-08-01 from https://service.elsevier.com/app/answers/detail/a_id/31662/supporthub/scopus/

Simpson, E. H. (1949). Measurement of diversity. *Natur*e, 163(4148), 688.

Skupien, S., & Rüffin, N. (2019). The geography of research funding: Semantics and beyond. *Journal of Studies in International Education*, 24(1), 24-38.

Vieira, E. S. (2022). International research collaboration in Africa: A bibliometric and thematic analysis. *Scientometrics*, 127(5), 2747-2772.

Waltman, L., Eck, N. J. v., Visser, M., et al. (2024, 2024-01-30). Introducing the Leiden ranking open edition. Retrieved 2024-08-01 from https://www.leidenmadtrics.nl/articles/introducing-the-leiden-ranking-open-edition

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744), 378-382.

Xu, F., Wu, L., & Evans, J. (2022). Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences*, 119(23), e2200927119.

Appendix

Table 5. Country (region) information.

Country full name	Country code	If African country (region)	Country full name	Country code	If African country (region)
Afghanistan	AF		Lesotho	LS	√
Aland Islands	AX		Liberia	LR	√
Albania	AL		Libyan Arab Jamahiriya (the)	LY	√
Algeria	DZ	√	Liechtenstein	LI	
American Samoa	AS		Lithuania	LT	
Andorra	AD		Luxembourg	LU	
Angola	AO	√	Macao	MO	
Anguilla	AI		Macedonia (the former Yugoslav Republic of)	MK	
Antarctica	AQ		Madagascar	MG	√
Antigua and Barbuda	AG		Malawi	MW	√
Argentina	AR		Malaysia	MY	
Armenia	AM		Maldives	MV	
Aruba	AW		Mali	ML	√
Australia	AU		Malta	MT	
Austria	AT		Marshall Islands (the)	MH	
Azerbaijan	AZ		Martinique	MQ	
Bahamas (The)	BS		Mauritania	MR	√
Bahrain	BH		Mauritius	MU	√
Bangladesh	BD		Mayotte	YT	√
Barbados	BB		Mexico	MX	
Belarus	BY		Micronesia (the Federated States of)	FM	
Belgium	BE		Moldova (the Republic of)	MD	
Belize	BZ		Monaco	MC	
Benin	BJ	√	Mongolia	MN	
Bermuda	BM		Montenegro	ME	
Bhutan	BT		Montserrat	MS	
Bolivia	BO		Morocco	MA	√
Bosnia and Herzegovina	BA		Mozambique	MZ	√
Botswana	BW	√	Myanmar	MM	
Bouvet Island	BV		Namibia	NA	√
Brazil	BR		Nauru	NR	
British Indian Ocean Territory (the)	IO		Nepal	NP	
Brunei Darussalam	BN		Netherlands (the)	NL	

Bulgaria	BG		Netherlands Antilles (the)	AN	
Burkina Faso	BF	√	New Caledonia	NC	
Burundi	BI	√	New Zealand	NZ	
Cambodia	KH		Nicaragua	NI	
Cameroon	CM	√	Niger (the)	NE	√
Canada	CA		Nigeria	NG	√
Cape Verde	CV	√	Niue	NU	
Cayman Islands (the)	KY		Norfolk Island	NF	
Central African Republic (the)	CF	√	Northern Mariana Islands (the)	MP	
Chad	TD	√	Norway	NO	
Chile	CL		Oman	OM	
China	CN		Pakistan	PK	
Christmas Island	CX		Palau	PW	
Cocos (Keeling) Islands (the)	CC		Palestinian Territory (the Occupied)	PS	
Colombia	CO		Panama	PA	
Comoros	KM	√	Papua New Guinea	PG	
Congo	CG	√	Paraguay	PY	
Congo (the Democratic Republic of the)	CD	√	Peru	PE	
Cook Islands (the)	CK		Philippines (the)	PH	
Costa Rica	CR		Pitcairn	PN	
Côte d'Ivoire	CI	√	Poland	PL	
Croatia	HR		Portugal	PT	
Cuba	CU		Puerto Rico	PR	
Cyprus	CY		Qatar	QA	
Czech Republic (the)	CZ		Réunion	RE	√
Denmark	DK		Romania	RO	
Djibouti	DJ	√	Russian Federation (the)	RU	
Dominica	DM		Rwanda	RW	√
Dominican Republic (the)	DO		Saint Helena	SH	√
Ecuador	EC		Saint Kitts and Nevis	KN	
Egypt	EG	√	Saint Lucia	LC	
El Salvador	SV		Saint Pierre and Miquelon	PM	
Equatorial Guinea	GQ	√	Saint Vincent and the Grenadines	VC	
Eritrea	ER	√	Samoa	WS	
Estonia	EE		San Marino	SM	
Ethiopia	ET	√	Sao Tome and Principe	ST	√
Falkland Islands (the) [Malvinas]	FK		Saudi Arabia	SA	
Faroe Islands (the)	FO		Senegal	SN	√

Fiji	FJ		Serbia	RS	
Finland	FI		Seychelles	SC	√
France	FR		Sierra Leone	SL	√
French Guiana	GF		Singapore	SG	
French Polynesia	PF		Slovakia	SK	
French Southern Territories (the)	TF		Slovenia	SI	
Gabon	GA	√	Solomon Islands (the)	SB	
Gambia (The)	GM	√	Somalia	SO	√
Georgia	GE		South Africa	ZA	√
Germany	DE		South Georgia and the South Sandwich Islands	GS	
Ghana	GH	√	Spain	ES	
Gibraltar	GI		Sri Lanka	LK	
Greece	GR		Sudan (the)	SD	√
Greenland	GL		Suriname	SR	
Grenada	GD		Svalbard and Jan Mayen	SJ	
Guadeloupe	GP		Swaziland	SZ	√
Guam	GU		Sweden	SE	
Guatemala	GT		Switzerland	CH	
Guernsey	GG		Syrian Arab Republic (the)	SY	
Guinea	GN	√	Taiwan (Province of China)	TW	
Guinea-Bissau	GW	√	Tajikistan	TJ	
Guyana	GY		Tanzania, United Republic of	TZ	√
Haiti	HT		Thailand	TH	
Heard Island and McDonald Islands	HM		Timor-Leste	TL	
Holy See (the) [Vatican City State]	VA		Togo	TG	√
Honduras	HN		Tokelau	TK	
Hong Kong	HK		Tonga	TO	
Hungary	HU		Trinidad and Tobago	TT	
Iceland	IS		Tunisia	TN	√
India	IN		Turkey	TR	
Indonesia	ID		Turkmenistan	TM	
Iran (the Islamic Republic of)	IR		Turks and Caicos Islands (the)	TC	
Iraq	IQ		Tuvalu	TV	
Ireland	IE		Uganda	UG	√
Isle of Man	IM		Ukraine	UA	
Israel	IL		United Arab Emirates (the)	AE	
Italy	IT		United Kingdom (the)	GB	

Jamaica	JM		United States (the)	US	
Japan	JP		United States Minor Outlying Islands (the)	UM	
Jersey	JE		Uruguay	UY	
Jordan	JO		Uzbekistan	UZ	
Kazakhstan	KZ		Vanuatu	VU	
Kenya	KE	√	Venezuela	VE	
Kiribati	KI		Viet Nam	VN	
Korea (the Democratic People's Republic of)	KP		Virgin Islands (British)	VG	
Korea (the Republic of)	KR		Virgin Islands (U.S.)	VI	
Kuwait	KW		Wallis and Futuna	WF	
Kyrgyzstan	KG		Western Sahara	EH	√
Lao People's Democratic Republic (the)	LA		Yemen	YE	
Latvia	LV		Zambia	ZM	√
Lebanon	LB		Zimbabwe	ZW	√

Table 6. Grouping of African countries.

Region	Country full name	Country code	Region	Country full name	Country code
Northern Africa	Algeria	DZ	Middle Africa	Angola	AO
	Egypt	EG		Cameroon	CM
	Libya	LY		Central African Republic	CF
	Morocco	MA		Chad	TD
	Sudan	SD		Congo	CG
	Tunisia	TN		Democratic Republic of the Congo	CD
	Western Sahara	EH		Equatorial Guinea	GQ
Eastern Africa	British Indian Ocean Territory	IO		Gabon	GA
	Burundi	BI		Sao Tome and Principe	ST
	Comoros	KM	Southern Africa	Botswana	BW
	Djibouti	DJ		Eswatini	SZ
	Eritrea	ER		Lesotho	LS
	Ethiopia	ET		Namibia	NA
	French Southern Territories	TF		South Africa	ZA
	Kenya	KE	Western Africa	Benin	BJ
	Madagascar	MG		Burkina Faso	BF
	Malawi	MW		Cabo Verde	CV
	Mauritius	MU		Côte d'Ivoire	CI
	Mayotte	YT		Gambia	GM
	Mozambique	MZ		Ghana	GH
	Réunion	RE		Guinea	GN

	Rwanda	RW		Guinea-Bissau	GW
	Seychelles	SC		Liberia	LR
	Somalia	SO		Mali	ML
	South Sudan	SS		Mauritania	MR
	Uganda	UG		Niger	NE
	United Republic of Tanzania	TZ		Nigeria	NG
	Zambia	ZM		Saint Helena	SH
	Zimbabwe	ZW		Senegal	SN
				Sierra Leone	SL
				Togo	TG

Table 7. Major funding institutions.

Full name	Abbreviation	Affiliated country
National Research Foundation	NRF	South Africa
Department for International Development	DFID	United Kindom
International Development Research Centre	IDRC	Canada
United States Agency for International Development	USAID	United States
Styrelsen för Internationellt Utvecklingssamarbete	Sida	Sweden
European Commission	EC	/
National Institutes of Health	NIH	United States
Bill and Melinda Gates Foundation	BMGF	United States
Economic and Social Research Council	ESRC	United Kingdom
World Bank Group	WBG	/
Addis Ababa University	AAU	Ethiopia
Consortium pour la recherche économique en Afrique	CREA	/
Covenant University	CoU	Nigeria
University of Cape Coast	UCC	Ghana
Water Research Commission	WRC	South Africa
University of Cape Town	UCT	South Africa
South African Medical Research Council	SAMRC	South Africa
University of Johannesburg	UJ	South Africa
Wellcome Trust	WT	United Kingdom
Social Sciences and Humanities Research Council of Canada	SSHRC	Canada
National Science Foundation	NSF	United States
Economic Research Forum	ERF	Egypt
Islamic Development Bank	IDB	/
Cairo University	CaU	Egypt
United Nations International Children's Emergency Fund	UNICEF	/
Ministry of Higher Education and Scientific Research	MHESR	Egypt
Agence Nationale de la Recherche	ANR	France
African Development Bank Group	ADBG	/
Centre for International Forestry Research	CIFOR	/
World Resources Institute	WRI	/

https://doi.org/10.51408/issi2025_027

Insights from Publication Timing: The Impact of Knowledge Features on the Disruptive Scores of Papers

Shan Huang[1], Jin Mao[2], Gang Li[3]

[1]*huangshan_gz16@whu.edu.cn*
Wuhan University, School of Information Management, No.299 Bayi Road, 430072 Wuhan (China)

[2]*maojin@whu.edu.cn*, [3]*ligang@whu.edu.cn1*
Wuhan University, Center for Studies of Information Resources, No.299 Bayi Road, 430072 Wuhan (China)

Abstract

Early identification of highly disruptive publications can improve resource allocation and accelerate scientific innovation. Many studies have examined the factors influencing paper disruption and methods for identifying them. However, most methods require at least three years after publication to assess the disruption of papers, which may not align with the demand of stakeholders for early identification of disruptive publications. Moreover, current studies often treat knowledge content as a supplement to citation-based approaches, while neglecting the intrinsic value of knowledge. To overcome these limitations, this study proposes six inherent knowledge features that can be recognized at the time of publication and try to reveal their function in shaping the disruption of papers. Specifically, we divide them as two categories, while "Knowledge linkage step," "Knowledge depth," and "Knowledge width" as structural features, "Knowledge age variance," "Knowledge age," and "Knowledge reuse" as attribute features. We then analyzed the relationship between these knowledge features and the disruption of papers using two datasets from biomedical science. The Golden Paper dataset includes 100 highly disruptive papers and 100 control papers; and the Large-scale dataset, which contains over 3 million papers. In the Golden Paper dataset, we balanced control variables using Entropy Balancing Matching (EBM), The empirical analysis shows that highly disruptive papers exhibit distinct characteristics. Compared to less disruptive papers at publication time, they contain more diverse and broadly distributed knowledge and rely on more recent knowledge Besides, they also exhibit lower knowledge reuse also revealed similar patterns, less depth and shorter linkages. The empirical analysis based on the Large-scale dataset also revealed similar patterns, knowledge age variance and knowledge width were positively correlated disruption scores, while higher knowledge age, knowledge reuse, and knowledge linkage step were associated with lower disruption scores. Additionally, we found that disruption scores in the Large-scale dataset showed a decreasing trend over the years, which may be related to opposing trends in knowledge feature distributions and their relationship with disruption scores. Specifically, the knowledge age, depth, reusability, and linkage steps of knowledge show a small upward trend over time. However, these features are negatively correlated with the disruption scores. Our study encourages the early identification of disruptive papers by revealing the relationship between knowledge features and disruption, offering insights for early prediction of disruptive papers in biomedical science.

Introduction

Disruptive scientific innovation is a key driver of paradigm shifts in modern science, which transcends disciplinary boundaries and reshapes scholars' understanding. According to Kuhn's (1962) theory of scientific revolutions, the evolution of science progresses through alternating phases of normal science and scientific revolution (Leibel & Bornmann, 2024). Normal science follows established paradigms, with innovation occurring gradually through the accumulation of knowledge. In contrast,

a scientific revolution disrupts existing paradigms, leading to major breakthroughs and steering science in new directions (Lin et al., 2022). After that, science returns to a new normal phase, waiting for the next scientific revolution. Scientific revolutions are often driven by disruptive innovations. Christensen (1997) introduced the concept of "disruptive innovation" in the context of marketing and described disruption as "the process by which a small company with few resources can successfully challenge the established firms. " In scientific publications, disruptive innovation represents a leap in the knowledge trajectory, probably leading to a shift in the knowledge paradigm (Funk & Owen-Smith, 2017; Leibel & Bornmann, 2024). Because these leaps may lead to substantial scientific advancements, publications characterized by high disruptive innovation are increasingly attracting the attention of scientists.

In response to the growing interest in highly disruptive papers, scholars have increasingly focused on developing accurate identification methods, most of which rely on citation network analysis. Disruption index (DI) and their variants, such as the Journal Disruption Index (JDI) and the Interdisciplinary Disruption Index (IDI), are typical citation-based methods (Funk & Owen-Smith, 2017; Jiang & Liu, 2023; Chen et al., 2024). After being cited by two highly impact papers published in *Nature*, the DI has become a representative method for identifying disruptive publications (Wu et al. 2019; Park et al. 2023). According to the concept of the Disruption Index (DI), a paper is considered disruptive if it tends to "replace" its foundational citations in subsequent research. The greater its deviation from previous citation patterns, the more disruptive it is considered to be (Bornmann et al., 2020; Wuestman et al., 2020). However, while the DI and its variants are widely used, studies have found that their accuracy is influenced by factors such as time window, citation inflation, and limited data coverage (Leibel & Bornmann, 2024; Petersen et al., 2024). Moreover, these methods fail to address the "Sleeping Beauty" problem, where disruptive papers may remain dormant for years before their value is recognized, limiting the speed of scientific evolution (Van Raan, 2004; Li & Ye, 2016; Hartley & Ho, 2017). These constraints demonstrate the need to reduce biases from citation and data that affect the disruption identification of publications. In addition, identifying highly disruptive papers before the public recognized their relevance is equally important.

Early detection of potentially highly disruptive papers plays a vital role in accelerating the evolution of science, particularly when such recognition occurs in the year of publication. Many highly disruptive papers show few visible signs in the early stages, and the information available is limited at these stages (Xu et al., 2022). Therefore, scientists have attempted to identify early predict factors of disruption by analysing paper features, with author-related and reference-related factors being the most representative. On one hand, the number of authors is negatively correlated with disruption, while teams with authors from monodisciplinary background or a higher proportion of young scientists tend to produce more disruptive outcomes (Wu et al., 2019; Liu et al., 2024). On the other hand, papers citing references from a single field tend to have lower disruption scores, while references from multiple disciplines may indicate interdisciplinary innovation, leading to higher disruption

scores (Chen et al., 2024; Yu et al.,2024). However, author and reference features primarily describe external aspect of a paper, while the knowledge concent of paper may carry more direct information of disruption.

Although the knowledge content of a paper has already been considered an inherent factor in publications (since it is fixed from the publication year), it is typically viewed as a supplement to complement citation-based measures of disruption rather than being observed as a subject independently. And these studies assume that all knowledge in a paper is equally important, with no difference. For example, Wang et al. (2023) proposed a measure of disruption score based on the impact of the knowledge created and used in academic papers on the trajectory of scientific evolution. Similarly, Lin et al. (2025) introduced the Disruptive Innovation Benchmark (DIB), which incorporates the scope of influence a paper has on subsequent publications based on knowledge trajectory measurement, to assess disruption. However, treating knowledge content as the main object of analysis rather than a supplement to citation-based measurement allows for the identification of key factors like the features of knowledge underlying disruptive publications that remain undetected by traditional citation-based methods.

Biomedical science provides an ideal domain for identifying the disruption of papers based on knowledge content, as it features a more structured and standardized knowledge organization compared to the other domains. It also benefits from the use of the well-established Medical Subject Headings (MeSH), which standardizes the knowledge in the publications. MeSH descriptors are organized in a hierarchical tree structure and updated annually (*"National Library of Medicine," n.d.*). MeSH terms closer to the root node represent broader knowledge, which covers more specific concepts, while those closer to the leaves denote more specific knowledge. This hierarchical structure can reveal hidden relationships and knowledge features that may be overlooked when treating all knowledge elements equally (Zheng et al., 2024b). Additionally, the annual updates managed by the NIH introduce new knowledge and adjust the positioning of existing knowledge in the tree to reflect developments in the biomedical sciences. Therefore, utilizing the MeSH tree structure from the year of publication to represent the knowledge framework is an ideal source for extracting the knowledge features of a paper.

This study proposes a series of knowledge features exhibited by papers at the time of publication and reveals the correlation between different knowledge features and the disruption of papers. We evaluated knowledge features in a publication from knowledge structure and knowledge attributes. The empirical analysis utilizes a Golden Paper dataset with highly disruptive papers and a Large-scale dataset with more than 3 million publications; both came from the biomedical sciences. The research questions are as follows:

RQ1: How do the knowledge feature of highly disruptive publications differ from others?

RQ2: Does the inherent features of knowledge in publications affect the disruption scores of the publications?

We contribute to the identification of scientific disruptive innovations in several ways. First, we used MeSH to distinguish the hierarchical structure and levels of

knowledge, which enhanced the understanding of the features of knowledge within papers. Second, we identified the influence of inherent knowledge features on the disruption scores of papers, revealing the relationship between them more clearly. This supports the possibility of identifying disruptive papers at the time of publication. Finally, by focusing on the inherent knowledge features of papers, we propose a new direction for the early prediction of disruptive innovations, offering a deeper understanding of the generation of highly disruptive papers in biomedical science.

Related work

Knowledge hierarchical structures and knowledge features

Scientific knowledge is inherently organized through hierarchical structures, which serve as foundational frameworks for categorizing and interpreting complex information (Clauset et al., 2008; Qian et al., 2020). Tree structure is a specialized form of hierarchical representation, where higher-level nodes represent broader conceptual scopes and lower-level nodes denote specialized subfields (Muchnik et al., 2007; Zheng et al., 2024b). Besides, the depth of a tree branch reflects the degree of specialization within a knowledge domain, measured by the number of sequential nodes (Geng et al., 2020). A branch with multiple nested nodes may indicate a well-developed research area, whereas shorter branches often correspond to emerging or less-explored knowledge topics. This structural property allows scientists to quantify knowledge features by analyzing positions of nodes. Recent studies have found that knowledge at higher levels in a hierarchy is usually more stable and connected across different fields because their position is nearer to the root node, while knowledge at lower levels has more potential for innovation (Yang et al., 2025).

In biomedical sciences, MeSH terms are organized hierarchically in the MeSH tree, including 16 main categories, and each category branches into subcategories, progressing from general to specific concepts. For instance, general categories like "Diseases" branch into specific conditions such as "Neurodegenerative Diseases" and further into granular terms like "Alzheimer's Disease" (*"National Library of Medicine," n.d.*). The hierarchical depth reflects conceptual specificity, enabling precise indexing of research themes. This structure allows researchers to analyze knowledge breadth (via parent terms) and depth (via child terms), while the introduction year of MeSH terms provides temporal insights into knowledge evolution (Zheng et al., 2024b). Therefore, the MeSH tree is suitable for the induction and analysis of knowledge features.

Scientists classify knowledge features into three main categories: structural features, attribute features, and temporal features. Structural features describe the overall configuration of knowledge, such as the range of topics covered and the level of specialization (Zheng et al., 2024b). For example, a paper with broad MeSH term coverage may exhibit greater knowledge breadth, while one with highly specific terms may show deeper specialization. Attribute features, on the other hand, focus on the intrinsic properties of knowledge and its position in a knowledge network (Yang et al., 2024). These features are often measured using complex network

metrics, which reveal how knowledge elements interact with each other (Wang et al., 2022; Yang & Hu, 2025). And temporal features capture the dynamic nature of knowledge, emphasizing how it evolves over time (Yang & Hu, 2025). All these features provide a comprehensive view of knowledge within scientific papers, offering insights into their potential impact and disruption.

Factors influencing the disruptive expression of papers

The concept of "disruption" is defined as the possibility to challenge existing paradigms and redirect research trajectories in publications (Funk & Owen-Smith, 2017). The more disruptive the paper, the more likely it is to change the existing research paradigm (Wei et al., 2023; Wuestman et al., 2020).

Recent studies on the disruption of publications have identified several key factors that shape their potential to challenge existing paradigms. These factors can be grouped into inherent features, which relate to the content of paper, and external factors, which concern the context in which the paper is published (He & Jing, 2024). Scholars have extensively studied the inherent features of authors and reference patterns. Papers authored by senior scientists often gain recognition more quickly but may be less disruptive, as they tend to their align with mainstream ideas. In contrast, work produced by early-career researchers or monodisciplinary teams tends to introduce novel perspectives, which is more likely to increase the disruptive potential in their research (Liu et al., 2024; Jiang et al., 2024). However, higher productivity among authors in a paper may be associated with lower levels of paper disruption (Li et al., 2024). Reference features also influence a paper's potential to be disruptive. Papers that cite older or foundation references tend to build upon established knowledge, whereas those citing recent and unconventional work are more likely to challenge existing paradigms. (Chen et al., 2024; Yu et al., 2024). Nevertheless, current studies often overlook knowledge-based features, especially the structural and attributive features of knowledge within papers. These features reflect the intrinsic organization of the knowledge of a paper and may provide important insights into the mechanisms of disruption, yet they have not been fully explored.

More importantly, these knowledge features are static and can be analyzed as soon as a paper is published, unlike post-publication indicators, which evolve over time and are influenced by external factors (Christensen et al., 2018). By focusing on these inherent knowledge features, scientists can identify potential disruption early, even in the publication year of the paper. Therefore, investigating the relationship between a paper's knowledge features at publication and its disruption is essential for advancing our understanding of scientific innovation and identifying highly disruptive papers in the earliest stage.

Methodology

Data collection

We collected two datasets with Medical Subject Heading (MeSH) terms for empirical analysis. The first one is 100 golden breakthrough papers published between 2013 and 2018 in biomedical science, as well as the corresponding control group papers. Golden papers come from a set of top journals in the field of biomedical science. First, we collected the golden papers from 2013-2018 in the top journals, including The New England Journal of Medicine (NEJM), The Journal of the American Medical Association (JAMA), and Cell. These journals publish about 10 highly disruptive papers each year in the form of news or electronic publications. Due to missing indexing on some pages, we manually collected 108 eligible papers, of which only 100 papers with more than 1 MeSH term as golden papers entered the dataset. Secondly, we collected 2,136 publications that were published in the same journal, year, volume, and issue as the golden papers, considering them as a potential control group. Then, a one-to-one random matching was conducted between the golden papers and the potential control papers, resulting in 100 matched papers. These selected papers were designated as the matched control group (low disruption) for comparison with the high-disruption group.

Another set of data is publications coming from the PubMed database, which was used to investigate how the knowledge features effect the disruption in a large-scale quantitative analysis. Large-scale dataset was retrieved from the prior works by Liang et al (2021), they built a dataset, which was expanded PubMed2020 baseline by adding citation data from Web of Science and NIH-OCC, providing biomedical science data and MeSH terms of over 30 million publications. We only retained publications with the number of MeSH terms more than 1 and with 10 or more references and cited literature for the study (Wang et al., 2023). Publications from 2015 onwards were removed because papers in the 5-year window at the time of data collection did not ensure the accuracy of the disruptive index measurement. These processes resulted in a final dataset of 3,590,997 publications as focal papers (FP) with publication years between 2001-2015 (Figure 1). include papers from 2001 onwards because the "MeSH tree" information showed in the MeSH browser is more completed after that year.

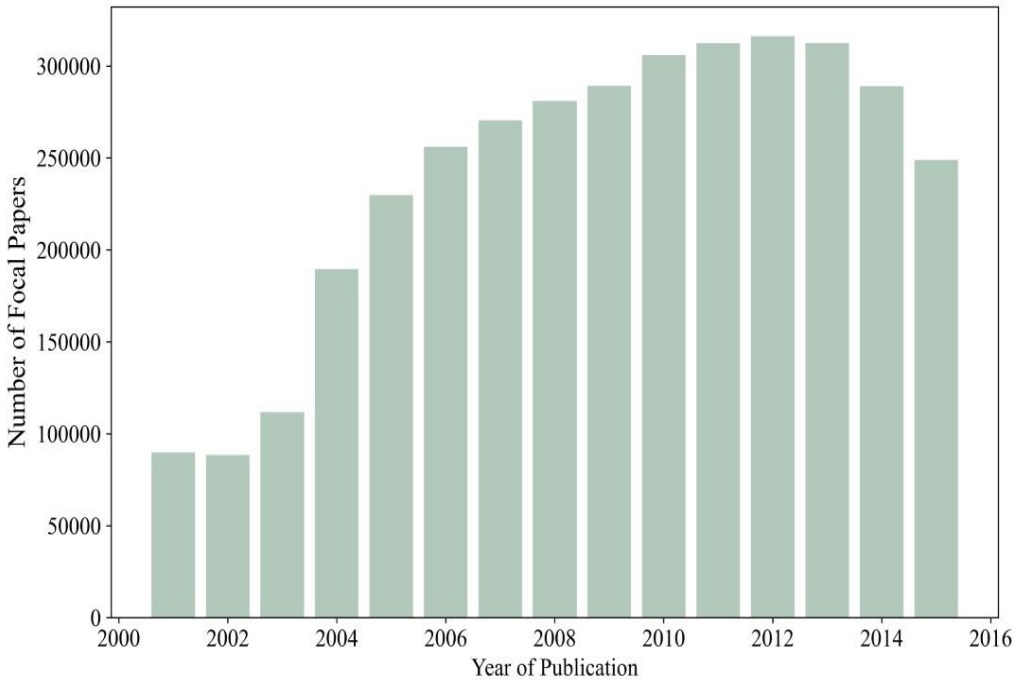


Figure 1. The distribution of the FPs in large-scale dataset from PubMed over years.

MeSH-based knowledge features

The MeSH Thesaurus, introduced by the U.S. National Library of Medicine (NLM), is organized into a hierarchical structure known as the MeSH tree. It serves as a standardized terminology system and provides comprehensive coverage of medical topics. Figure 2(a) shows a part of the whole MeSH tree. Besides, a MeSH term can appear at multiple levels within the hierarchy. The MeSH terms positioned closer to the end of the hierarchy represented more specific knowledge descriptions.

As the most authoritative content thesaurus list in the biomedical sciences, the MeSH tree is regularly updated each year to reflect the latest advances in medical knowledge and technology. Updates to the MeSH tree help scientists stay informed about the latest knowledge structure as well as the dynamic changes in knowledge hierarchical structure. In order to determine the attributes of MeSH terms at the time of publication, we retrieved the corresponding MeSH tree for each paper's publication year from the MeSH browsers (*"National Library of Medicine," n.d.*). In this way, we can calculate all the knowledge features of each paper at their publication year.

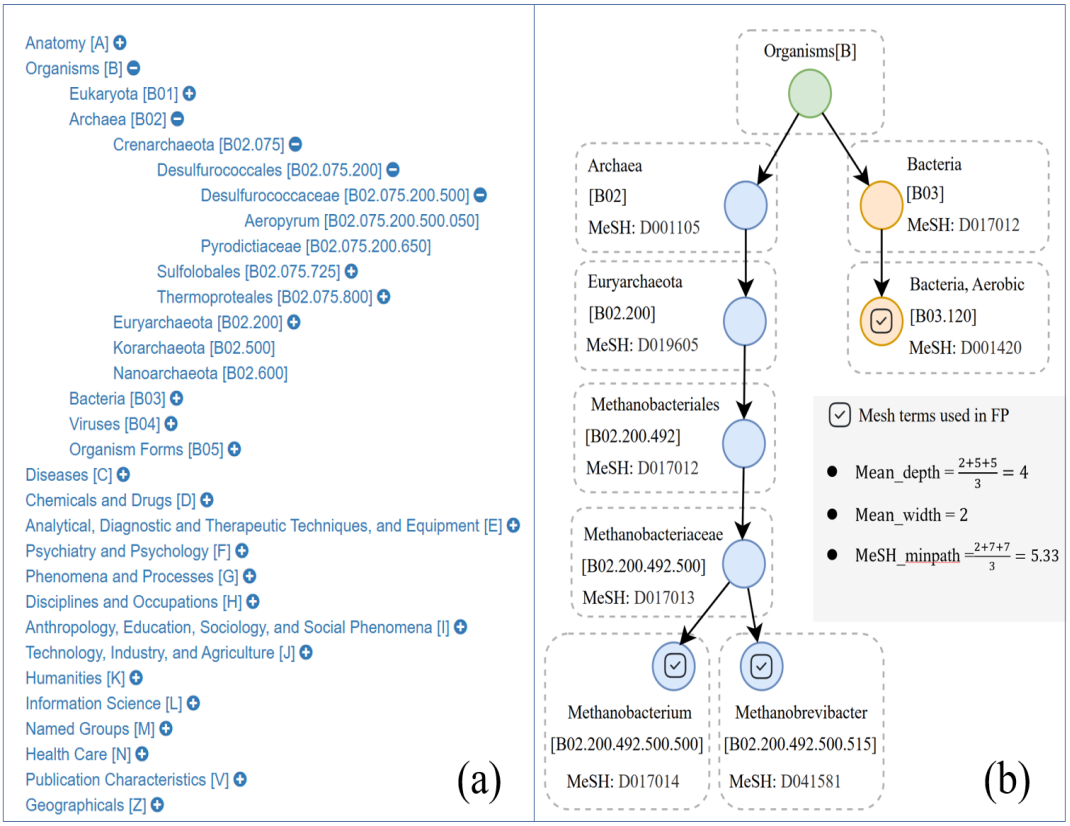


Figure 2. Examples for MeSH tree (a) and structure features calculation of a focal paper (b).

We propose six knowledge features based on MeSH thesaurus and MeSH tree hierarchy, and divide these features into two categories according to their sources. The structure features are derived from the position of knowledge in the MeSH tree, including knowledge depth, knowledge width and knowledge linkage step. The attribute features describe the properties of knowledge, including knowledge age, knowledge age variance and knowledge reuse.

The hierarchical structure of the MeSH tree shares similarities with the evolution of knowledge diffusion patterns. Rowlands (2002) introduced the concept of Data Knowledge Diffusion Breadth (DKDB) to analyze the diffusion range of knowledge. Goldman (2014) highlighted that node at the initial stage of a diffusion path tend to occupy more central positions in the network than terminal nodes. Drawing on the features of diffusion breadth and intensity, we propose two structural features of the MeSH tree: Mean depth and knowledge width. We hypothesize that the position of the knowledge used in a paper, as represented in the MeSH tree, reflects the organizational structure of the research content. Figure 2(b) provides an example of the calculation.

Knowledge depth: Represents the specificity of the research content. It is calculated as the average hierarchical level of all MeSH terms used in a focal paper (Eq. 1). Where M_d is the depth of MeSH term, n is the number of mesh terms in FP.

$$Mean_{depth} = \frac{1}{n}\sum_{m=1}^{n} M_d \quad (1)$$

Knowledge width: The average number of independent knowledge domains covered by all MeSH terms in the paper, reflecting the knowledge coverage of the study. Here, the second-level nodes of the MeSH tree (e.g., [B01]) are used as independent knowledge domains (Eq. 2). Where M_c is the domain in which term m is located, $Count\ domain$ only keeps the number of domains that are not duplicated, and n is the total number of terms m in FP.

$$Mean\_width = Count\ domain(\sum_{m=1}^{n} M_c) \quad (2)$$

Besides, the tightness of the connection of knowledge in the paper in the mesh tree can represent the degree of knowledge aggregation, which can be represented by the average shortest connection step between knowledge in the paper.

Knowledge linkage step: By pairing the MeSH terms in the paper, the shortest path between each pair in the MeSH tree is calculated, and the average of these shortest paths represents the tightness of knowledge connections in the paper (Eq. 3). Where (p_m, p_{m+1}) is the link step and n is the total number of MeSH terms m in FP. For example, [B02.200.492.500.500] and [B02.200.492.500.515] in Figure 2(b) have the same upper node, and their connection step is only 2.

$$MeSHmin_{path} = \frac{2*\sum_{m}^{n}\sum_{m+1}^{n} shortest\ path(p_m,p_{m+1})}{n(n-1)} \quad (3)$$

The strategic usage of both recent and diverse knowledge sources, which can collectively facilitate breakthroughs in science and technology (Mukherjee et al., 2017). Inspiring by researchers' findings that impactful research leverages both recent and temporally diverse knowledge, we adapt knowledge age and knowledge age variance as two of the attribute features in publications are defined as follows:

Knowledge age: The temporal gap between the first appearance t_0 of a MeSH term and its use in the focal paper which published in year t. The mean age of knowledge of a paper is measured as the average of the ages of all the MeSH terms used in the papers (Eq. 4). Where m denotes a MeSH term used in the focal paper, n is the number of MeSH terms.

$$Mean_{year} = \frac{1}{n}\sum_{m=1}^{n}\big(t(m) - t_0(m)\big) \quad (4)$$

Knowledge age variance: The dispersion of reference ages, which is represented the temporal diversity of knowledge. The value of this feature will be expressed by calculating the variance of the age of knowledge in the focal paper (Eq. 5). Where a_m is the age of each MeSH term, $\bar{a}$ denotes the average knowledge age in FP.

$$Sd_{year} = \frac{1}{n}\sum_{m=1}^{n}(a_m - \bar{a})^2 \quad (5)$$

Besides, knowledge reuse is also a feature of external attributes, which used to measure and characterize the prevalence and generality of MeSH terms in a paper. The annual average reuse count of each MeSH term was calculated from its first appearance to the publication year of the paper. Higher reuse count indicates stronger acceptance in science, showing that this MeSH term is more widely used The knowledge reuse of a single paper is measured by the average reuse count of all its MeSH terms, which is used to portray the level of acceptance of the knowledge contained in the paper at the time of publication (Eq. 6).

$$MeSH_{reuse} = \frac{1}{n}\sum_{m=1}^{n}\frac{N_m}{t_{pub} - t_{first} + 1} \tag{6}$$

Where N_m is the total number of occurrences of MeSH m, t_{pub} and t_{first} represent the year of publication of FP and the year of m's first appearance, respectively.

Matching analyses

We employ Entropy Balancing Matching (EBM) method and Mann-Whitney U test to observe whether the difference of each knowledge feature existing or not between high disruption publications and the normal disruption papers. EBM is applicable for group-level matching, which is more suitable for balancing the confounding variables in our study. Meanwhile, Mann-Whitney U test is used to assess the significance of the differences of knowledge features.

Entropy Balancing Matching (EBM) was introduced by Hainmueller (2012), which utilized changes in information entropy to match treatment and control groups at the level of confounding factors. This approach aims to balance the distribution of the control variables between the high disruption and ordinary papers, reducing the effect of confounders on the dependent variables to effectively compare the different performance of the independent variables between the two groups. In this study, we used EBM to ensure confounding balance for papers in the treatment group and control group, so that knowledge features were comparable between the treatment and control groups. This adjustment allows for more clearly revealing of how knowledge features may affect the disruptive expression of research.

EBM involves three key steps: assigning weights to control group individuals to balance covariates between the treatment and control groups, calculating the information entropy increment between the treatment group and the weighted control group, and selecting the result with the minimal entropy increment as the counterfactual estimate (Hainmueller, 2012; Zheng et al., 2024a).

$$\hat{E}[P(0)|Disruption = 1] = \frac{\sum_{\{i|Disruption = 0\}} P_i w_i}{\sum_{\{i|Disruption = 0\}} w_i} \tag{7}$$

The Eq. 7 demonstrates the computation of the weighted control group estimate, which serves as the counterfactual outcome for the group of highly disruptive papers. The left-hand side of the equation represents the expected counterfactual value for each high disruptive paper, assuming it were less disruptive paper. P_i denotes the

observed outcome of each paper i in the control group, while w_i indicates the weight assigned to individual i after entropy balancing.

To evaluate the significance of differences in each knowledge feature between the treatment and control groups after EBM, we employed the Mann-Whitney U test. This non-parametric method is particularly suited for small sample sizes and data that do not follow a normal distribution. The steps of the test include:

(1) Hypothesis formulation: Setting null hypothesis H_0 that there is no significant difference between high and ordinary disruption papers; alternative hypothesis H_1 that there is a significant difference between the two groups of papers.

(2) Ranking and summation: All observations from the two groups (highly and less disruptive papers) are pooled and ranked together. The sum of rankings for each group is calculated separately, denoted as $SumD_0$ (less disruptive papers) and $SumD_1$ (highly disruptive papers)

(3) U-statistics test: The U-statistic for each group is computed using the follow Eq. 8 Where $SumD_k$ (k = 0 or 1) denotes the sum of ranking for one group, n denotes the number of observations in the group. The smaller U-value between the two groups is selected and compared to the critical value at a significance level (p = 0.05). If U-test < Up, we reject H_0 and confirm that there is a significant difference between high disruption and ordinary disruption papers, and vice versa.

$$U_{test} = \min\left(SumD_{k=1} - \frac{n(n-1)}{2}, SumD_{k=0} - \frac{n(n-1)}{2}\right) \quad (8)$$

The above steps are looped 6 times to obtain results for all knowledge features. In other words, knowledge features passing the test are considered to have significant effect on the disruption scores of papers, while those failing the test are excluded from further analysis.

Measuring disruption score

To determine the disruption scores of papers, we adopted two methods depended on the characteristics of the datasets. For the smaller dataset of Golden papers, we assigned fixed disruption scores: Golden papers were assigned a score of 1, representing high disruption, while ordinary papers were assigned a score of 0. And for the large-scale dataset, we applied an indicator-based approach to measure disruption scores. Although the indicators proposed by Wu et al (2019), which relies on the citation network of focal papers, has been widely acknowledged, its scope is limited to document-level analysis. Wang et al (2023) introduced disruption indicators that considers shifts in knowledge flow to optimised previous studies and validated them in biomedical datasets, incorporating the role of focal papers' knowledge content. Furthermore, such a series of disruption indicators were later expanded to WOS dataset by Tong et al (2024).

We focused on the impact of knowledge features on disruption scores in this study, and employing the ED index (ED) proposed by Wang et al. (2023) to calculate focal papers' disruption scores is more suitable. Figure 3 provides the content and patterns

to be observed when measuring this indicator. In the citation network related to the focal paper, nodes are represented by different shapes and shades of gray (diamonds for references and focal papers, while circles and squares denotes different types of citing papers separately). Knowledge elements are distributed across this network, and categorized into six types based on their frequency and position, represented as triangles. Wang et al (2023) used individual MeSH terms and their combinations as knowledge elements to measure disruption scores (ED_ent and ED_rels) separately. Therefore, we focused on mED_ent for main analysis and used ED_rels for robustness checks to ensure the reliable results and minimize bias.

The *ED* index measures the disruption scores of a focal paper by analyzing the flow and transformation of knowledge elements within its citation network. To account for the effect of citation inflation, the index incorporates a weighting parameter m as proposed by Funk and Owen-Smith (2017). The *ED* index consists of two components: the deviation of the focal paper's knowledge elements from its references ED_b and the extent to which the focal paper's new knowledge is reinforced by its citing papers $ED_{a,t}$, as shown in Eq. 9 and Eq. 10. Where N denotes the number of papers that share at least one MeSH term and citing FP, and n_{bi}, n_{bj}, n_{ai}, n_{an}, n_{aj} and n_{ak} represent the number of knowledge elements of the corresponding type, respectively. By setting β as a parameter, the ED index exhibits different behaviours under varying parameter values. Since no specific component is emphasized in this study, the ED index is calculated as the average of the two components (β = 0.5), as shown in Eq. 11.

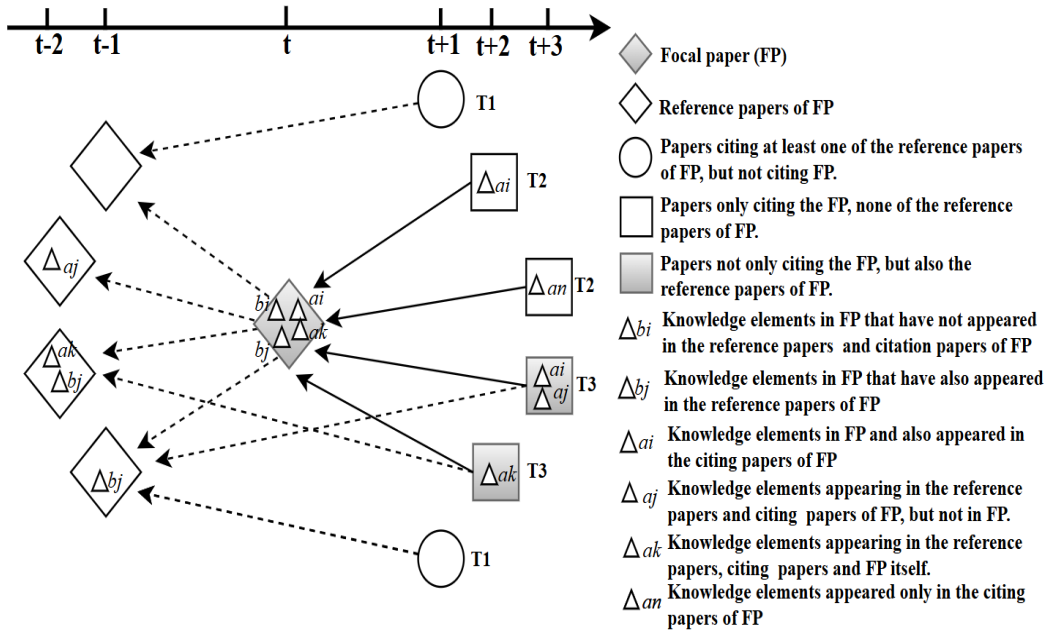


Figure 3. Illustration of the citation pattern with knowledge elements related to FP (Wu et al., 2019; Wang et al., 2023).

$$ED_b = \frac{n_{bi} - n_{bj}}{n_{bi} + n_{bj}} \quad (9)$$

$$ED_{a,t} = \frac{1}{N}\sum_{c=1}^{N} \frac{n_{ai} + n_{an} - n_{aj} - n_{ak}}{n_{ai} + n_{an} + n_{aj} + n_{ak}} \quad (10)$$

$$ED_t = \beta ED_b + (1-\beta) ED_{a,t} \quad (11)$$

Regression models for knowledge features and disruption score of publications

Disruption score was used as the dependent variable in our study, which was measured by ED_ent, with individual MeSH terms serving as the knowledge elements. And six types of knowledge features were employed as independent variables in the regression models. Besides, several factors except knowledge features may affect the disruption of publications, which should be controlled in the regression models. Previous studies revealed that the characteristics of metadata in papers, especially the number of authors and references, were fully correlated with disruption scores (Wu et al., 2019; Petersen et al., 2024). Similarly, maintaining the number of MeSH terms at a consistent level may help enhance comparability between papers. Therefore, we selected these factors as control variables.

Number of authors: The number of authors represents the team size of publications. Recent studies show that small teams tend to produce more disruptive publications and software compared to large teams (Wu et al., 2019). However, large teams with high organizational diversity may also generate high disruptive outcomes (Yoo et al., 2024).

Number of references: A longer reference list is one of the key factors of citation inflation, leading to the density of citation networks of publications, which may distort the calculation of a publication's disruptive score (Petersen et al., 2024).

Number of MeSH terms: Citation inflation is usually accompanied by an increase in the amount of knowledge in the publications. Controlling the number of MeSH terms contributes to reducing the influence of knowledge inflation.

Furthermore, the publication years were adjusted to minimize potential influence. Table 1 exhibits the details of all types of variables.

By considering the dependent variable as a continuous variable, we employed the Ordinary Least Squares (OLS) regression model to investigate the relationship between knowledge features and disruption scores in publications. Eq. 12 shows the basic regression model.

$$Disruption_i = \alpha_0 + \alpha_1 Mean\_year_i + \alpha_2 Sd\_year_i + \alpha_3 Mean\_depth_i + \alpha_4 Mean\_width_i + \alpha_5 MeSH\_reuse_i + \alpha_6 MeSHmin\_path_i + \alpha_7 Control_i + PY_i + \varepsilon_1 \quad (12)$$

Where $Disruption$ denotes the disruption score of the FP i, Mean_year , Sd_year, $Mean\_depth$, $Mean\_width$, $MeSH\_reuse$ and $MeSHmin\_path$ denote the knowledge features of FP i separately, $Control$ contains all the control variables, PY_i is the year of publication fixed effects, and $\varepsilon_1$is the error term.

Table 1. The list of variables used in OLS regression models.

Variables	*Symbol of variables*	*Description of variables*
Diruption scores	ED_ent	The disruption scores of FP calculated by using individual MeSH terms as knowledge element.
Knowledge age variance	Sd_year	Age variance of knowledge used by FPs.
Knowledge age	Mean_year	Average age of knowledge used by FP.
Knowledge reuse	MeSH_reuse	Average number of times the knowledge in the FP appeared in the prior publications up to the years when FP was published.
Knowledge linkage step	MeSHmin_path	The average step size when the knowledge used by FP is pairwise connected.
Knowledge depth	Deep_mean	The average depth of the knowledge used by FP in the Mesh tree hierarchy.
Knowledge width	Wide_mean	The number of branches covered in the Mesh tree by the knowledge used by FP.
Number of MeSH	Len_MeSH	The number of individual MeSH terms of a FP.
Reference number	Ref_num	The number of references of a FP.
Number of authors	AuthorNum	The number of authors of a FP.
Publication year	Pub_year	The publication year of a FP.

Result

Knowledge features of highly disruptive publications

We used the EBM approach to balance the differences based on selected control variables between highly disruptive papers and less disruptive papers in the Golden Paper dataset. Less disruptive papers (control group) were matched to highly disruptive papers (treatment group) using these variables. After matching, a balance test checked if the matching worked well. The results showed that the control variables for retracted articles between highly disruptive papers and ordinary papers were balanced, as displayed in Figure 4, which made the comparison more reliable and reduced the impact of control variables on the results.

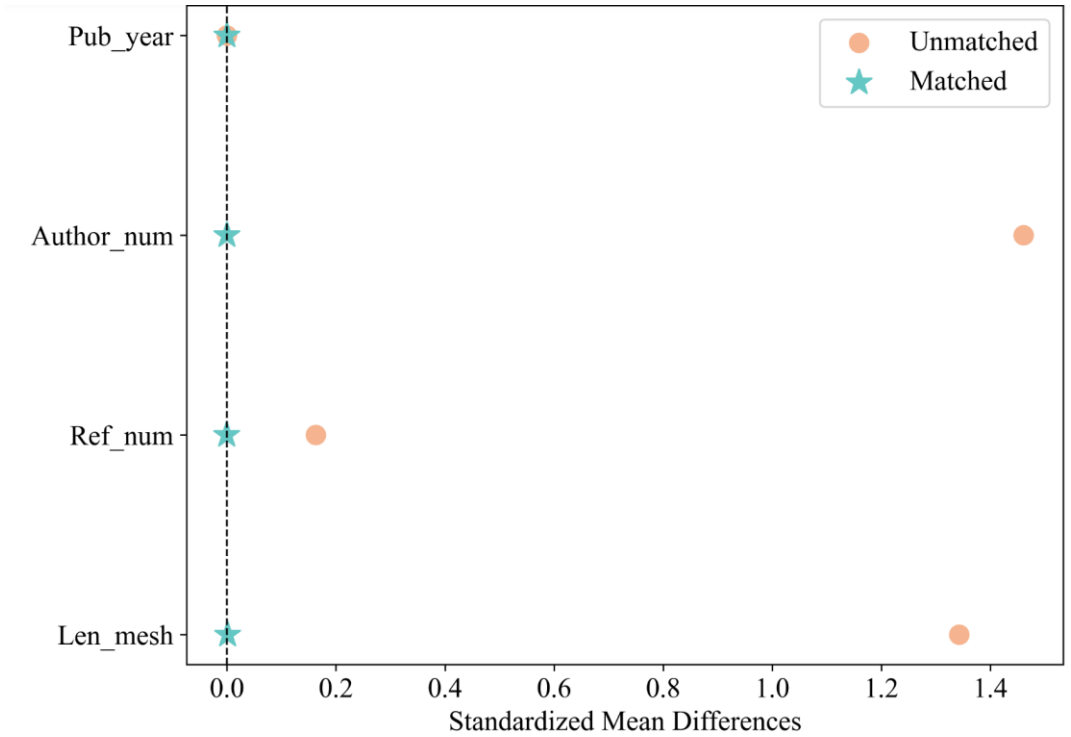


Figure 4. Standardized mean differences of control variables for papers between groups of highly and less disruptive papers before and after EBM.

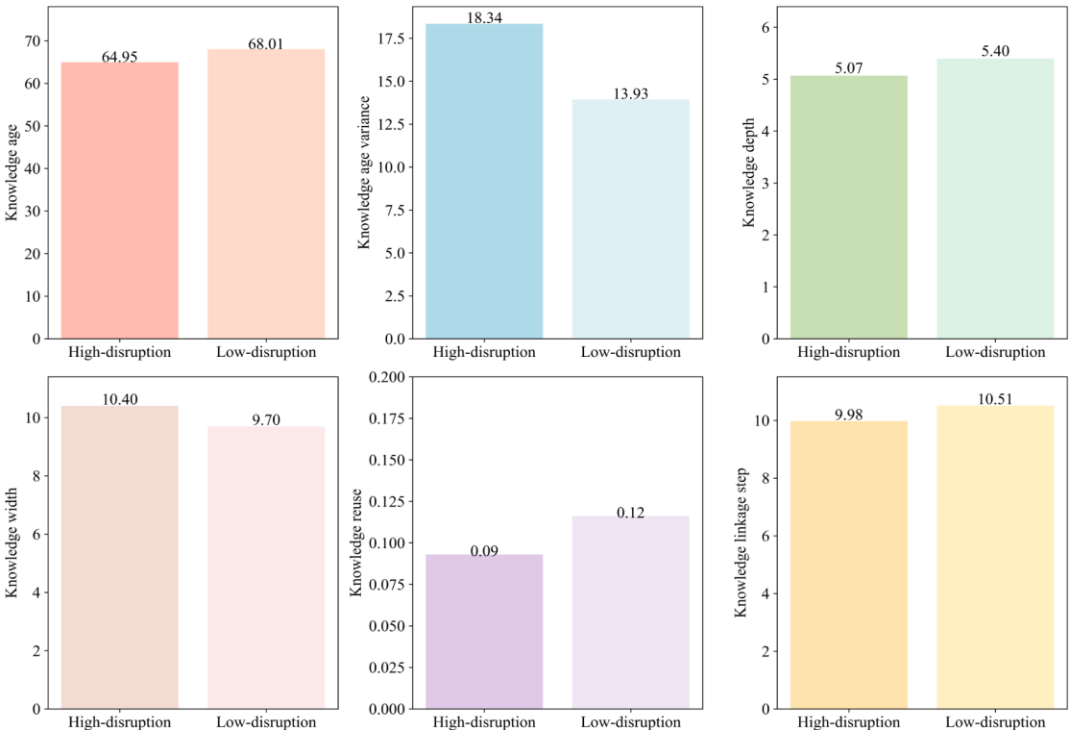


Figure 5. Knowledge features differences of papers between groups of highly and less disruptive papers after EBM.

Following the balance test, we calculated the average scores of the six knowledge features using balancing weights derived from the EBM process. Figure 5 shows the differences in knowledge features between the two groups. The golden papers demonstrated higher knowledge age variance (18.34 vs. 13.93) and a greater average knowledge width (10.40 vs. 9.70), indicating that highly disruptive papers tend to incorporate more diverse and broadly distributed knowledge under EBM balance. Conversely, the knowledge age, knowledge depth, knowledge reuse, and knowledge linkage step were all lower for highly disruptive papers compared to normal disruption papers. These results suggest that highly disruptive papers are characterized by younger knowledge, lower reuse at the time of publication, as well as less knowledge depth, and shorter knowledge linkages.

Furthermore, we employed the Mann-Whitney U test for each feature to assess whether a statistically significant difference exists between highly disruptive papers and ordinary papers. The null hypothesis *(H0)* assumed no significant difference between the two groups, while the alternative hypothesis *(H1)* proposed a significant difference. The results are summarized in Table 2. The *Z-score* representing the standardized U statistic, a positive *Z-score* indicates that highly disruptive papers exhibit higher values for the knowledge feature compared to ordinary papers. When the absolute value of the *Z-score* exceeds 3.29, the difference is statistically significant at the 0.001 level, choosing the hypothesis *(H1)*. Obviously, all six knowledge features were found to show significant differences, suggesting that highly disruptive papers are characterized by distinct knowledge features at the time of publication when compared to ordinary papers.

Table 2. The results of the Mann-Whitney U test for the difference in each knowledge feature.

Knowledge feature	*Z-value*	*P-value*	*Hypothesis selection*
Mean_year	-4.833	P<0.001	H1
Sd_year	6.152	P<0.001	H1
Deep_mean	-3.366	P<0.001	H1
Wide_mean	5.186	P<0.001	H1
MeSH_reuse	-8.366	P<0.001	H1
MeSHmin_path	-4.933	P<0.001	H1

The trends of knowledge features and disruption scores for all the biomedical publications

We observe the trends of six knowledge features of the publications in the Large-scale dataset over years, as shown in Figure 6. The trend of the knowledge width demonstrates volatility over the years, but stabilizes at relatively low values after 2009.In contrast, the values of the other five features show a continuous upward trend over years. In terms of the attribute features of knowledge, papers tend to use more established and older knowledge, with a diversity in the age distribution of knowledge used. Regarding the structure features of knowledge, the low mean width

indicates that the papers cover a narrow and specialized range of topics. In contrast, the increasing depth of knowledge suggests that studies increasingly focus on more specific knowledge located deeper within the MeSH hierarchy. In addition, the increasing trend in the mean pathway of knowledge link demonstrates that the knowledge in the papers may span across different branches, with longer path connection.

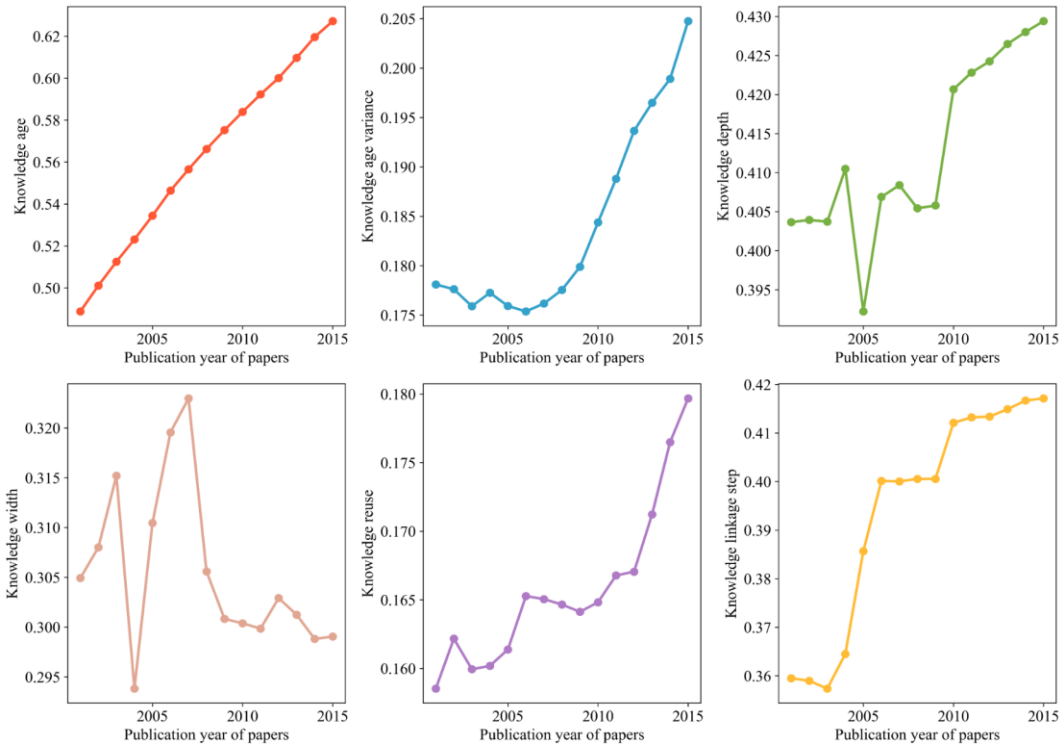


Figure 6. Trends of knowledge features for the publications of Large-scale dataset over years.

The distribution of disruption scores shows a slow decline over years (Figure 7), which is resemble to the results reporting in Nature by Park et al (2023), including decreases in the upper and lower bounds (after removing outliers) and the median value. This may indicate that more recent papers rely increasingly on established knowledge, limiting the possibility to change the evolutionary trajectory of knowledge within the biomedical science (Wang et al., 2023).

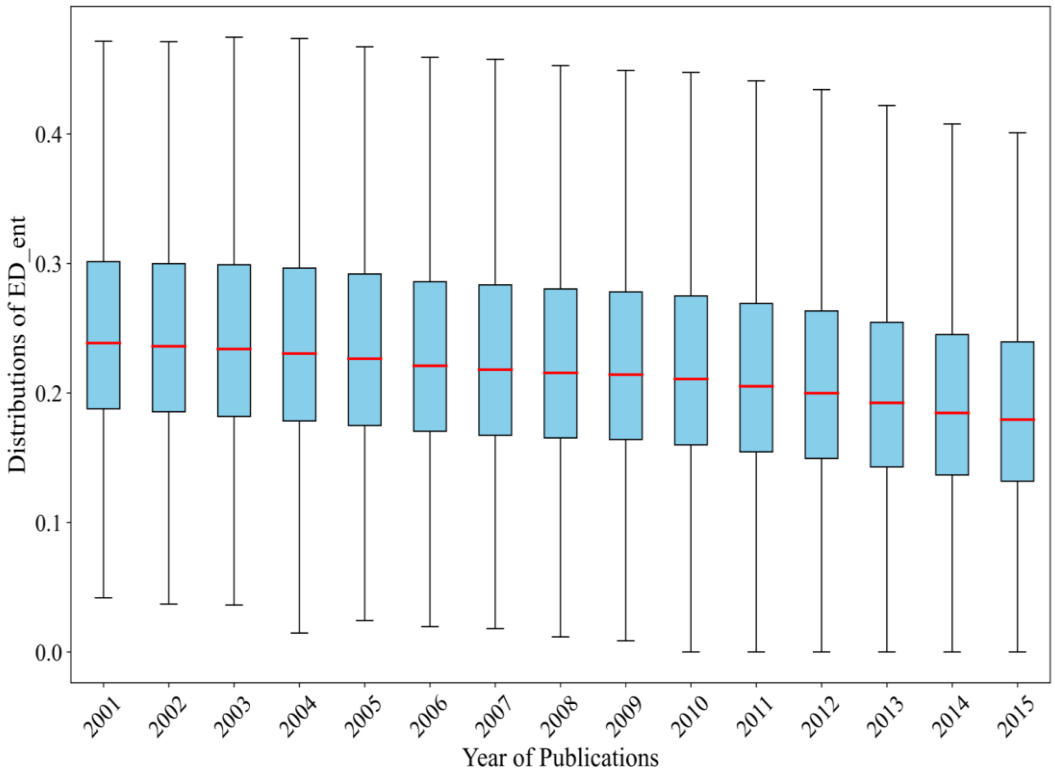


Figure 7. The distribution of disruption scores in the publications of large-scale dataset and their trends over years.

Regression analysis

We analyzed the relationship between knowledge features and disruption scores of publications using regression models. Table 3 reported the results. Model 1 only contains the control variables and disruption score. Model 2 and 3 show the effects of structural features and attribute features on disruption scores of publications, respectively. Model 4 uses all variables. The independent variables display consistent patterns across the models, highlighting the stability of these relationships. Specifically, higher knowledge age variance is significantly associated with higher disruption scores, and similar positive correlation results are found for the knowledge width. However, higher knowledge age, knowledge reuse and knowledge linkage step were negatively associated with disruption scores.

Table 3. Estimated relationships between knowledge features and ED_ent disruption scores in Large-scale dataset.

Disruption	*ED_ent*			
	Control (1)	*Structure (2)*	*Attribute (3)*	*All features (4)*
Sd_year			0.0195***	0.0324***
			(0.0006)	(0.0005)
Mean_year			-0.0260***	-0.0835***
			(0.0007)	(0.0009)
MeSH_reuse			-0.2082***	-0.1576***
			(0.0004)	(0.0005)
Pmidmin_path		-0.1926***		-0.1170***
		(0.0008)		(0.0009)
Deep_mean		-0.1286***		-0.1046***
		(0.0008)		(0.0008)
Wide_mean		0.0213***		0.0251***
		(0.0006)		(0.0006)
Len_MeSH	-0.0318***	-0.0362***	-0.0721***	-0.0876***
	(0.0004)	(0.0005)	(0.0004)	(0.0005)
Ref_num	-3.3562***	-3.3208***	-3.2662***	-3.3273***
	(0.0051)	(0.0051)	(0.0051)	(0.0051)
AuthorNum	-1.8771***	-1.1085***	-2.0184***	-1.6520***
	(0.0253)	(0.0250)	(0.0246)	(0.0246)
Pub_year	YES	YES	YES	YES
const	7.1876***	5.0782***	6.4454***	4.1094***
	(0.0255)	(0.0260)	(0.0292)	(0.0313)
Obs.	3590997	3590997	3590997	3590997
F-test	137860.4306	96632.6677	115018.6653	84475.4704
R^2	0.1331	0.1585	0.1831	0.1904

Note: Robust standard errors in parentheses. *p < 0.05, **p < 0.01, ***p < 0.001.

To ensure the robustness of the relationships between the six features and disruption scores, we tested alternative methods for calculating the dependent variable and adjusted the regression approach (Table 4). First, we replaced individual MeSH terms with MeSH combinations as the knowledge elements for dependent variable measurement, both of which were provided by Wang et al (2023). Model 1 and 2 show the relationship results when ED_ent (measuring by individual MeSH term) and ED_rels (measuring by MeSH combination) are used as dependent variables, respectively. Second, we employed Stepwise Regression model (SR) to replace OLS regression model and randomly selected 80% of the sample from the Large-scale dataset as the test data, with the results shown in Model 3 and 4. SR not only identifies the suitable set of predictors but also addresses multicollinearity issues. All of the results confirm that the correlations between knowledge features and disruption scores remain significantly robust across all checking cases.

Table 4. Robustness check based on different disruption scores, and Stepwise Regression models.

Disruption	*OLS Regression model*		*Stepwise Regression model*	
	ED_ent (1)	*ED_rels (2)*	*ED_ent (3)*	*ED_rels (4)*
Sd_year	0.0324***	0.1432***	0.0313***	0.1463***
	(0.0005)	(0.0008)	(0.0006)	(0.0009)
Mean_year	-0.0835***	-0.3335***	-0.0894***	-0.2825***
	(0.0009)	(0.0013)	(0.0009)	(0.0014)
MeSH_reuse	-0.1576***	-0.4230***	-0.1571***	-0.4460***
	(0.0005)	(0.0007)	(0.0006)	(0.0009)
Pmidmin_path	-0.1170***	-0.1897***	-0.1056***	-0.1395***
	(0.0009)	(0.0013)	(0.0011)	(0.0016)
Deep_mean	-0.1046***	-0.2067***	-0.1187***	-0.2376***
	(0.0008)	(0.0012)	(0.001)	(0.0015)
Wide_mean	0.0251***	0.0947***	0.0242***	0.0785***
	(0.0006)	(0.0008)	(0.0006)	(0.0010)
Len_MeSH	-0.0876***	-0.0647***	-0.0861***	-0.0462***
	(0.0005)	(0.0008)	(0.0006)	(0.0009)
Ref_num	-3.3273***	-3.6381***	-3.3197***	-3.4991***
	(0.0051)	(0.0076)	(0.0057)	(0.0085)
AuthorNum	-1.6520***	-2.1774***	-1.6168***	-2.1241***
	(0.0246)	(0.0369)	(0.0285)	(0.0426)
Pub_year	YES	YES	YES	YES
const	4.1094***	0.5940***	0.4546***	0.9280***
	(0.0313)	(0.0469)	(0.0007)	(0.0010)
Obs.	3590997	3590997	2872797	2872797
F-test	84475.4704	114885.0076	67037.5705	91791.0351
R^2	0.1904	0.2424	0.1892	0.2421

Note: Robust standard errors in parentheses. *p < 0.05, **p < 0.01, ***p < 0.001.

Conclusion and discussion

Early identification of publications with disruptive potential can significantly enhance the strategic allocation of scientific resources, fostering more efficient research system. We proposed six knowledge features based on the contents of publications at the time of their publication, including structural features and attribute features. This study conducted an in-depth analysis of the inherent knowledge features of papers to explore how these features differ from highly disruptive papers and less disruptive papers from the Golden Paper dataset. Furthermore, we confirm the critical role of these knowledge features in disruption scores by the analysing their relationship in a large-scale dataset of biomedical science. The findings quantitatively demonstrate significant correlations between knowledge features and disruption scores, offering a new perspective for identifying disruptive papers at an early stage.

High vs. Low Disruption of Papers: Differences in knowledge features at publication time

We shift the focus of identifying disruptive publications from citation networks to the features of knowledge at the year of paper publication, which provides a new perspective for early identification of disruptive papers. Empirical results reveal significant differences in knowledge features across papers at the time of publication between the groups of highly disruptive and less disruptive papers. Furthermore, a large-scale data analyses confirm associations between these features and disruption scores.

Specifically, highly disruptive papers exhibit distinct knowledge features compared to less disruptive papers at the time of publication in the Golden Paper dataset. They are associated with greater diversity in knowledge age, lower average knowledge age, and less reuse of knowledge. Moreover, they tend to demonstrate lower knowledge depth and shorter path lengths, and broader knowledge coverage. Similarly, in the Large-scale dataset of biomedical science, knowledge features such as knowledge age variance and knowledge width are positively correlated with disruption scores. In contrast, the knowledge age, knowledge depth, and the distance of knowledge connections exhibit significant negative correlations.

Our empirical findings indicate that it may be possible to identify highly disruptive study at the time of publication, rather than several years later as traditionally measured approaches (e.g., using DI1, DI5) (Wu et al., 2019; Funk & Owen-Smith, 2017). Unlike methods that depend on citation networks, we emphasize the inherent knowledge features at the publication time of papers. Specifically, we use MeSH terms to represent the knowledge content of each paper and calculate knowledge features. Our findings highlight the value of knowledge features in assessing scientific contributions. In addition, this approach effectively addresses limitations in citation-based approaches, such as citation inflation and time delay, which often bias disruption measurements (Petersen et al., 2019; Petersen et al., 2024). While our findings exhibit only a correlation between knowledge features and disruption scores of publications, this study may offer a useful perspective for understanding how highly disruptive works emerge.

Misaligned knowledge utilization may correlate with declining of disruption scores

We observed a consistent decline in the disruption scores of papers over the years in large-scale biomedical datasets. This trend aligns with the findings of Park et al (2023), who reported a similar decrease in disruption across 45 million documents. Park et al (2023) attributed this decline to a narrowing use of prior knowledge, where researchers increasing rely on well-established knowledge rather than exploring unconventional knowledge. This finding suggests a growing tendency to build on the "shoulders of giants", instead of venturing into less-charted research areas. Our study supports this perspective from the viewpoint of knowledge utilization.

In other word, the declining trend in the disruption scores of papers may be partially explained by changes in how knowledge is utilized in the publications. Specifically, we reveal that the knowledge features such as knowledge age, depth, reusability, and linkage distance have shown a slight upward trend over years, indicating that more recent publications tend to depend on older, more reusable, more specific knowledge,

with longer distances between knowledge connections. However, these features are negatively correlated with disruption scores. The opposing trends between the actual distribution of knowledge feature and the traits typically found in highly disruptive papers indicate some misalignments in knowledge utilization strategies. In other words, our findings reveal a significant difference in knowledge utilization features in most of the recent papers from the knowledge features observed in highly disruptive papers. These findings provide new insights on how shifts in knowledge utilizations might associated with the broader decline in disruption scores.

Limitations and future work

Although we have obverse that knowledge features are significantly affect the disruption score of publications, several limitations remain. First, while our analysis reveals the significant correlations between knowledge features and disruption score, we have not yet systematically evaluated the effectiveness of knowledge features in predicting highly disruptive papers at the early stage, which remains a key direction for future researches. Second, our empirical analysis focused on the biomedical science. Although we include both Golden papers and Large-scale dataset validation, the findings have not been extended to broader scientific disciplines. Lastly, due to current limitations in algorithms and computational resources, we are unable to dynamically collect the features of individual knowledge elements within complex knowledge networks at the time of publication. Although recognizing the potential importance of these features, we could not fully incorporate them in this work. Future studies may explore how to capture the evolution of knowledge and construct network-based modelling address this gap.

Acknowledgments

This study was funded by the National Natural Science Foundation of China (NSFC) Grant Nos. 71921002 and 72474159.

References

Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. Quantitative Science Studies, 1(3), 1242-1259.

Chen, S., Guo, Y., Ding, A. S., & Song, Y. (2024). Is interdisciplinarity more likely to produce novel or disruptive research?. Scientometrics, 1-18.

Christensen, C. M. (1997). *The innovator's dilemma: when new technologies cause great firms to fail*. Harvard Business Review Press.

Christensen, C. M., McDonald, R., Altman, E. J., & Palmer, J. E. (2018). Disruptive innovation: An intellectual history and directions for future research. *Journal of management studies*, *55*(7), 1043-1078.

Clauset, A., Moore, C., & Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. *Nature*, *453*(7191), 98-101.

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management science*, *63*(3), 791-817.

Geng, Z., Chen, G., Han, Y., Lu, G., & Li, F. (2020). Semantic relation extraction using sequential and tree-structured LSTM with attention. *Information Sciences*, *509*, 183-192.

Goldman, A. W. (2014). Conceptualizing the interdisciplinary diffusion and evolution of emerging fields: The case of systems biology. *Journal of informetrics*, *8*(1), 43-58.
Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, *20*(1), 25-46.
Hartley, J., & Ho, Y. S. (2017). Who woke the sleep beauties in psychology?. *Scientometrics*, *112*, 1065-1068.
He, Y., & Jing, J. (2024). Intellectual structure of disruptive innovation: a bibliometric analysis and systematic review. *Journal of Organizational Change Management*, *37*(6), 1382-1402.
Jiang, H., Zhou, J., Ding, Y., & Zeng, A. (2024). Overcoming recognition delays in disruptive research: The impact of team size, familiarity, and reputation. *Journal of Informetrics*, *18*(4), 101549.
Jiang, Y., & Liu, X. (2023). A construction and empirical research of the journal disruption index based on open citation data. *Scientometrics*, *128*(7), 3935-3958.
Kuhn, T. S. (1962). The structure of scientific revolutions. *University of Chicago Press*.
Leibel, C., & Bornmann, L. (2024). What do we know about the disruption index in scientometrics? An overview of the literature. *Scientometrics*, *129*(1), 601-639.
Li, H., Tessone, C. J., & Zeng, A. (2024). Productive scientists are associated with lower disruption in scientific publishing. *Proceedings of the National Academy of Sciences*, *121*(21), e2322462121.
Li, J., & Ye, F. Y. (2016). Distinguishing sleep beauties in science. Scientometrics, 108, 821-828.
Liang, Z., Mao, J., Lu, K., & Li, G. (2021). Finding citations for PubMed: a large-scale comparison between five freely available bibliographic data sources. *Scientometrics*, 126, 9519-9542.
Lin, R.H., Li, Y.L., Ji, Z., X, Q.Q., & Chen, X.Y. (2025). Quantifying the degree of scientific innovation breakthrough: Considering knowledge trajectory change and impact. *Information Processing & Management*, *62*(1), 103933.
Lin, Y., Evans, J. A., & Wu, L. (2022). New directions in science emerge from disconnection and discord. Journal of Informetrics, 16(1), 101234.
Liu, X., Bu, Y., Li, M., & Li, J. (2024). Monodisciplinary collaboration disrupts science more than multidisciplinary collaboration. Journal of the Association for Information Science and Technology, 75(1), 59-78.
Muchnik, L., Itzhack, R., Solomon, S., & Louzoun, Y. (2007). Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, *76*(1), 016106.
Mukherjee, S., Romero, D. M., Jones, B., & Uzzi, B. (2017). The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot. *Science advances*, *3*(4), e1601315.
National Library of Medicine. (n.d.). *MeSH (Medical Subject Headings)*. Retrieved November 14, 2024, from https://www.nlm.nih.gov/databases/download/mesh.html
Park, M., Leahey, E., & Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, *613*, 138–144
Petersen, A. M., Arroyave, F., & Pammolli, F. (2024). The disruption index is biased by citation inflation. *Quantitative Science Studies*, *5*(4), 936-953.
Petersen, A. M., Pan, R. K., Pammolli, F., & Fortunato, S. (2019). Methods to account for citation inflation in research evaluation. *Research Policy*, *48*(7), 1855-1865.

Qian, Y., Liu, Y., & Sheng, Q. Z. (2020). Understanding hierarchical structural evolution in a scientific discipline: A case study of artificial intelligence. *Journal of Informetrics*, *14*(3), 101047.

Rowlands, I. (2002, April). Journal diffusion factors: a new approach to measuring research influence. In *Aslib Proceedings* (Vol. 54, No. 2, pp. 77-84). MCB UP Ltd.

Tong, T., Wang, W., & Fred, Y. Y. (2024). A complement to the novel disruption indicator based on knowledge entities. *Journal of Informetrics*, *18*(2), 101524.

Van Raan, A. F. (2004). Sleep beauties in science. *Scientometrics*, *59*, 467-472.

Wang, S., Ma, Y., Mao, J., Bai, Y., Liang, Z., & Li, G. (2023). Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. *Journal of the Association for Information Science and Technology*, *74*(2), 150-167.

Wang, X., He, J., Huang, H., & Wang, H. (2022). MatrixSim: A new method for detecting the evolution paths of research topics. *Journal of Informetrics*, *16*(4), 101343.

Wei, C., Li, J., & Shi, D. (2023). Quantifying revolutionary discoveries: Evidence from Nobel prize-winning papers. *Information Processing & Management*, *60*(3), 103252.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378-382.

Wuestman, M., Hoekman, J., & Frenken, K. (2020). A typology of scientific breakthroughs. Quantitative Science Studies, 1(3), 1203-1222.

Xu, H., Luo, R., Winnink, J., Wang, C., & Elahi, E. (2022). A methodology for identifying breakthrough topics using structural entropy. *Information Processing & Management*, *59*(2), 102862.

Yang, J., & Hu, J. (2025). Scientific knowledge role transition prediction from a knowledge hierarchical structure perspective. *Journal of Informetrics*, *19*(1), 101612.

Yang, J., Liu, Z., & Huang, Y. (2024). From informal to formal: scientific knowledge role transition prediction. *Scientometrics*, *129*(8), 4909-4935.

Yoo, H. S., Jung, Y. L., Lee, J. Y., & Lee, C. (2024). The interaction of inter-organizational diversity and team size, and the scientific impact of papers. *Information Processing & Management*, *61*(6), 103851.

Yu, Q., Li, X., Ma, D., Zhang, L., Chen, K., Xue, Q., & Zhang, Q. (2024). Interdisciplinary hierarchical diversity driving disruption. Scientometrics, 129(12), 7833-7849.

Zheng, E. T., Fang, Z., & Fu, H. Z. (2024a). Is gold open access helpful for academic purification? A causal inference analysis based on retracted articles in biochemistry. *Information Processing & Management*, *61*(3), 103640.

Zheng, Z., Ma, Y., Ba, Z., & Pei, L. (2024b). Tree knowledge structure for better insight: Capturing biomedical science-technology knowledge linkage with MeSH. *Journal of Informetrics*, *18*(4), 101568.

https://doi.org/10.51408/issi2025_043

Interdisciplinarity and Artificial Intelligence: A Two-Dimensional Analysis of Diversity and Cohesion

Sisi Li[1], Yuxian Liu[2], Ronald Rousseau[3], Hongrui Yang[4], Sanfa Cai[5], Xiaoyan Zhang[6]

[1]*2231429@tongji.edu.cn*
Tongji University, Institute of Higher Education, Siping Road 1239, 200092 Shanghai (China)

[2]*yxliu@tongji.edu.cn*
Tongji University, Institute of Higher Education, Siping Road 1239, 200092 Shanghai (China)
Tongji University, Tongji University Library, Siping Road 1239, 200092 Shanghai (China)

[3]*ronald.rousseau@kuleuven.be*
University of Antwerp, Faculty of Social Sciences (Belgium)
Department of MSI, Centre for R&D Monitoring (ECOOM) (Belgium)

[4]*2330024@tongji.edu.cn*
Tongji University, Institute of Higher Education, Siping Road 1239, 200092 Shanghai (China)

[5]*csf@tongji.edu.cn*
Tongji University, School of political science & international relations, Siping Road 1239, 200092 Shanghai (China)

[6]*xyzhang@tongji.edu.cn*
Tongji University, School of Life Sciences and Technology, Siping Road 1239, Shanghai 200092 (China)

Abstract

Nowadays Artificial Intelligence (AI) has increasingly become the core technology leading a new round of scientific and technological revolution and industrial transformation. AI is reaching out to all disciplines and breaking the border of disciplines. Hence many universities are using AI as a strategic tool to empower traditional disciplines. Tongji University in Shanghai is one of these universities. We use Tongji University's publications on AI as a case to study the interdisciplinarity that led by AI studies. We collect the AI publications which have Tongji University in the address. The publications are sliced into five four-year slices and classified according to the categories of their published journals. We then calculate the categories' diversity and cohesion based on the categories of the references of these AI publications. We plot the categories in a Cartesian coordinate system with diversity as one coordinate and cohesion as the other. The results indicate that categories involved in AI studies have gradually evolved towards high diversity and high cohesion. Initially, the categories "Computer Science, Artificial Intelligence" were in the Quadrant III in the period of 2004-2007, then directly settled in the Quadrant I in the period of 2008-2012, maintaining high cohesion and high diversity ever since. This trend is closely tied to the development of computational science and deep learning technologies. Additionally, categories such as "Computer Science, Interdisciplinary Applications", "Engineering, Electrical & Electronic", "Biochemical Research Methods", "Biotechnology & Applied Microbiology", "Mathematics & Computational Biology" and "Operations Research & Management Science" were already classified as high cohesion-high diversity categories in the early stages (before 2016), demonstrating strong interdisciplinary integration capabilities. "Neuroscience," show a significant increase in its academic influence after

2016. The history of these disciplines reveals the crucial role that AI has played in driving the advancement of the concerned categories.

Introduction

Currently, the fourth wave of the technological revolution, represented by artificial intelligence (AI), is rapidly evolving. As a comprehensive and interdisciplinary field, AI's innovations rely heavily on cross-disciplinary integration with various disciplines. Previous studies have shown that AI can discover new knowledge and generate new technologies through the transplantation of theories, method exchange, and object transfer, further breaking down disciplinary barriers and expanding disciplinary boundaries. This process leads to a higher-level, more integrated collaboration between AI and other disciplines, spawning emerging fields and producing disruptive, original breakthroughs (Wu, 2019). The advent of AI has promoted the convergence between different disciplines, bringing about profound transformations in the development of those fields (Cai, Wang, & Shen, 2019). Therefore, assessing the degree of disciplinary integration in AI-related interdisciplinary research fields, tracking the evolving trends of cross-disciplinary convergence driven by AI, are of paramount importance for understanding its transformative impact on a disciplinary system.

Interdisciplinary research is generally seen as a source of creativity and innovativeness (Dogan and Pahre, 1990). The citation relationships among scientific literature represent the integration and diffusion of knowledge, indicating the direction of knowledge flow (Liu & Rousseau, 2010). Liu, Rafols and Rousseau (2012) introduce a general framework for the analysis of knowledge integration and diffusion using bibliometric data. They proposed to capture the characteristic of interdisciplinarity by the calculation of diversity and coherence from bibliometrics data. Measuring how particular articles integrate research fields based on the assignation of the journals they cite to WoS Subject Categories, is one of the approaches for assessing interdisciplinary (Porter, 2006).

Rafols and Meyer (2010) state that diversity and coherence are the two basic notions for the study of interdisciplinarity. Diversity refers to the breadth in categories used and consists of three basic concepts: variety, balance and disparity (Stirling, 2007). Coherence refers to the extent to which different elements in the research (categories or topics) are interrelated. The notion of diversity puts the emphasis on how different the incorporated knowledge is, while the notion of coherence emphasizes how different bodies of research are consistently articulated and form a meaningful constellation. In this sense, an increase in diversity reflects the divergence of knowledge integration and diffusion, whereas an increase in coherence reflects their convergence (Rousseau et al., 2019). Measuring interdisciplinarity from the perspective of coherence helps to elucidate the distinct roles of each discipline within the overall network and to identify the dominant disciplines driving the interdisciplinary integration process.

The studies provide valuable insights for examining the diversity and cohesion of every discipline involved in AI-related interdisciplinary fields from a two-dimensional perspective.

In the face of the rapidly evolving field of AI, universities are actively exploring new models for interdisciplinary integration, such as the "AI+" model of discipline construction, to empower and transform the development of their academic fields through AI. In 2019, Tongji University became one of the first universities in the country to be authorized to establish an undergraduate program in AI, and it also took the lead in developing the discipline of "Smart Science and Technology". In 2021, Tongji University was approved to develop an interdisciplinary doctoral program in "Smart Science and Technology", initiating research in AI-driven interdisciplinary fields. On May 16, 2024, Tongji University released the "Action Plan for AI-Driven Discipline Innovation and Development (2024-2027)," launching eight core tasks to strengthen the development of AI-related disciplines. Taking Tongji University as an example, measuring the diversity and cohesion of discipline of AI-related interdisciplinary fields is crucial for understanding the evolution of AI-driven discipline integration.

Data

To study the trends and situation of disciplinary integration caused by AI, this paper selects Tongji University as a case study and collects all published AI papers by Tongji University. The Web of Science core collection database is used as the data source. We refer to the search terms employed in the 2018 China Artificial Intelligence Development Report published by the Tsinghua University Science and Technology Policy Research Center, combined with expert opinions. The search strategy is as follows: (TS=("artificial intelligence" OR "machine learning" OR "natural language processing" OR "computer vision" OR "facial recognition" OR "image recognition" OR "speech recognition" OR "semantic search" OR "semantic web" OR "text analytics" OR “virtual assistance" OR “visual search" OR" predictive analytics" OR" intelligent system" OR "Deep Learning" OR "Robotics" OR "Autonomous Systems" OR "Human-Computer Interaction" OR "ChatGPT")) AND AD=(Tongji), no time frame was set, and the "Affiliation" filter was applied to select "TONGJI UNIVERSITY". This search strategy resulted in the retrieval of 3,839 articles retrieved, conducted on October 3, 2024. After deduplication, 3,367 valid articles were retained for subsequent subject categories mapping.

Based on the Journal Citation Reports (JCR) from the WoS, which provides the subject category information of each journal (a total of 254 WoS categories), we mapped each reference's corresponding journal to one or more WoS categories. A total of 157,761 reference records were mapped to categories. After data cleaning and mapping the references' categories, references with fewer than three subject categories were removed (Diego, Puay, & Rafols, 2014). As a result, 2,783 papers were selected as the sample for this study (spanning the period from 2004 to 2023). Dataset Availability: https://zenodo.org/records/15220666.

The data were preprocessed illustrated in Figure 1. Each publication was mapped to its categories of the journal that the publication was published. Each publication’s references were also mapped onto the categories of the references’ journals. The publications with same category were conglomerated, together with their references

and their categories, forming a map from the category of publications and its corresponding references' categories.

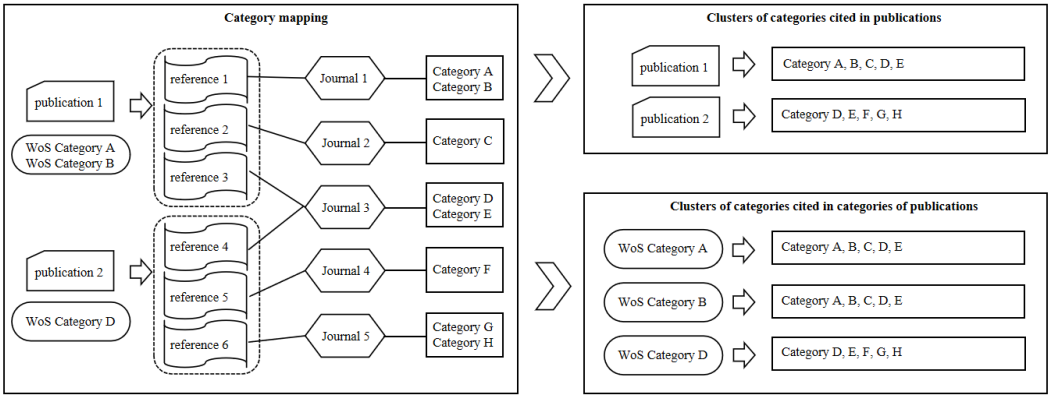


Figure 1. The mapping of the categories of publications and their references' categories.

The dataset was sliced into five four-year slices. Table 1 lists for every period the number of publications, the number of categories over which the publications are distributed, and the number of categories over which the references are distributed. The mapping in Figure 1 was also based on sliced dataset.

Table 1. Statistics of publications of Tongji University's artificial intelligence (AI) research field from 2004 to 2023.

Table	*Number of publications*	*Number of the categories the publications distributed*	*Number of the categories the references distributed*
2004-2007	17	21	45
2008-2011	47	35	98
2012-2015	103	42	140
2016-2019	450	102	205
2020-2023	2,166	171	237
2004-2023(total)	2,783	179	239

Methods

In this study, we adopt the conceptual framework proposed in (Rafols & Meyer, 2010) to measure the interdisciplinary of the categories related to AI research. Rafols and Meyer (2010) focused on measuring interdisciplinary of individual articles. Liu, Rafols & Rousseau (2012) enlarge this framework to a set of related articles. In this study, we measure the interdisciplinary of categories. the category is composed by a set of AI-related publications. So, we slightly modify the methods.

We use the distribution of references' categories as analytical unit to calculate the diversity. However, we have to conglomerate all publications in the same category

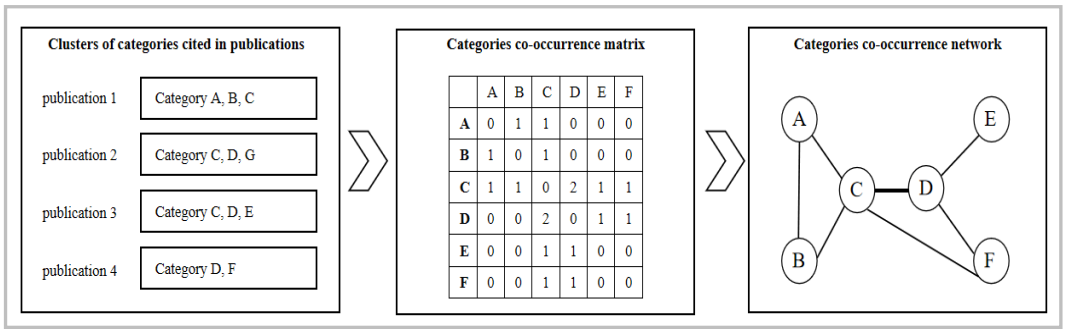
together to form a map from the category of AI publications to its references' categories. We calculate the category' diversity based on the conglomerate of the publications in same category in each time slice. Each publication is categorized into one or more categories by Web of Science. When conglomerating a publication which has several categories, the publication and its references' categories are conglomerated into all the categories the publication belongs.

We use Rao-Stirling indicator to measure the diversity of categories. The specific formula for the Rao-Stirling indicator is as follows:

$$\text{Diversity} = \text{Rao} - \text{Stirling} = \sum_{i,j(i \neq j)} (d_{ij})^{\alpha} \cdot (p_i \cdot p_j)^{\beta}{}_i$$

Here, d_{ij} denotes the dissimilarity between category i and category j. The dissimilarity of categories is calculated based on the inter-category co-membership of the journals in Web of Science, which was proposed by Liu (2018) to construct global backbone of science. p_i and p_j denote the proportions of the total number of items under study in category i and category j, respectively. Finally, α and β are parameters that adjust the importance given to small distances (α) and weights (β). In case one lacks empirical reasons to adjust α and β, they are often taken as being equal to 1.

We also construct a categories co-occurrence network in each time slice. Figure 2 shows how we constructed the network.



	A	B	C	D	E	F
A	0	1	1	0	0	0
B	1	0	1	0	0	0
C	1	1	0	2	1	1
D	0	0	2	0	1	1
E	0	0	1	1	0	0
F	0	0	1	1	0	0

Figure 2. The process of constructing a categories co-occurrence network.

Since we focus on the role of a category in integrating different categories related to AI, we do not go to measure the whole structural consistency of the AI publications network. A category can be considered to reside in an important mediating position in a network if it is on a path between many other categories. Leydesdorff (2007) suggested that betweenness centrality can be used as a measure of interdisciplinarity at the journal level. Rafols et al.(2012) developed intermediation as a framework, complementary to the diversity-cohesion framework. Zhang et al (2020) review the measurement for the cohesion of interdisciplinary research. Betweenness was acknowledged as a valid indicator to measure the cohesion of a single-node in the network.

In this paper, we chose the betweenness centrality of nodes to represent the degree of cohesion of categories.

Betweenness centrality is used to measure the role of each category as a bridge in the flow of information between categories. The formula of betweenness centrality is as follows:

$$\text{Cohesion} = \text{Betweenness Centrality} = \frac{\sum_{jk(j\neq k)} g_{jk}(i)}{g_{jk}}$$

Here, g_{jk} denotes the total number of shortest paths between two nodes; $g_{jk}(i)$ indicates the number of shortest paths between two nodes that pass-through node i.

Principles of Cartesian Coordinate System Plotting

Cohesion and diversity are the two dimensions we used to describe the interdisciplinary of the AI related categories. We establish a Cartesian coordinate system with diversity as one coordinate and cohesion as the other and plot the categories in the system according to our calculation.

To ensure comparability, all the data are standardized. In this process, each data point−representing cohesion or diversity for a given stage−is transformed relative to the overall dataset's mean. After standardization, values above the average are represented by positive values, while values below the average are represented by negative values. Hence the origin in the coordinate system represents the mean value for both cohesion and diversity across the entire 20-year period.

The primary advantage of this standardization process is that it enables the comparison of different stages in a uniform manner, with all data transformed onto a common scale.

Classification of Fields of Categories

Leydesdorff (2016) distinguished all categories of WOS into 5 broad categories or 18 fields. The five broad categories include: Social Sciences & Psychology, Engineering & Mathematics, Medicine, Physics & Chemistry, and Biology. The eighteen fields include: Social Sciences, Computer Science and Engineering, Medicine, Psychology, Environmental Sciences, Chemistry & Applied Physics, Biomedicine, Health Care, Engineering, Agriculture and Food, Management, Biology, Chemistry, Infectious diseases, Physics, Pharmacology, Environmental Engineering, Medicine & Others.

Drawing on Leidesdorff's classification of 18 fields, we reorganize the 254 categories in WoS into more generalized and conceptually coherent domains. The specific steps are as follows:

(1) First, the correspondence between the 227 WoS subject categories and the 18 major domains, as organized by Leydesdorff (2016), was adopted;

(2) A new field "Literature, History, Arts, and Philosophy"—was added, resulting in a final total of 19 major fields of categories;

(3) For the remaining 27 subject categories not covered in Leidesdorff's classification system, we assigned them to appropriate field based on expert judgment.

Result

Base on the publications of AI research field of Tongji University from 2004 to 2023, we calculate the indicators of the categories' diversity and cohesion and plot these categories in a Cartesian coordinate system. Below, we will analyze the two-dimensional feature of categories' cohesion and diversity for the periods 2004-2007, 2008-2011, 2012-2015, 2016-2019, and 2020-2023.

The two-dimensional feature of AI research field during 2004-2007

Figure 3 shows the two-dimensional feature analysis chart of the categories in the AI research field at Tongji University during 2004-2007, and Table 2 displays the corresponding classification table of the categories' two-dimensional features. A total of 21 Web of Science categories of publications were involved during this stage. The study also statistically analysed the categories with betweenness centrality of 0. To avoid excessive redundancy of data points in the scatter plot, these categories were not presented in the two-dimensional feature analysis chart.

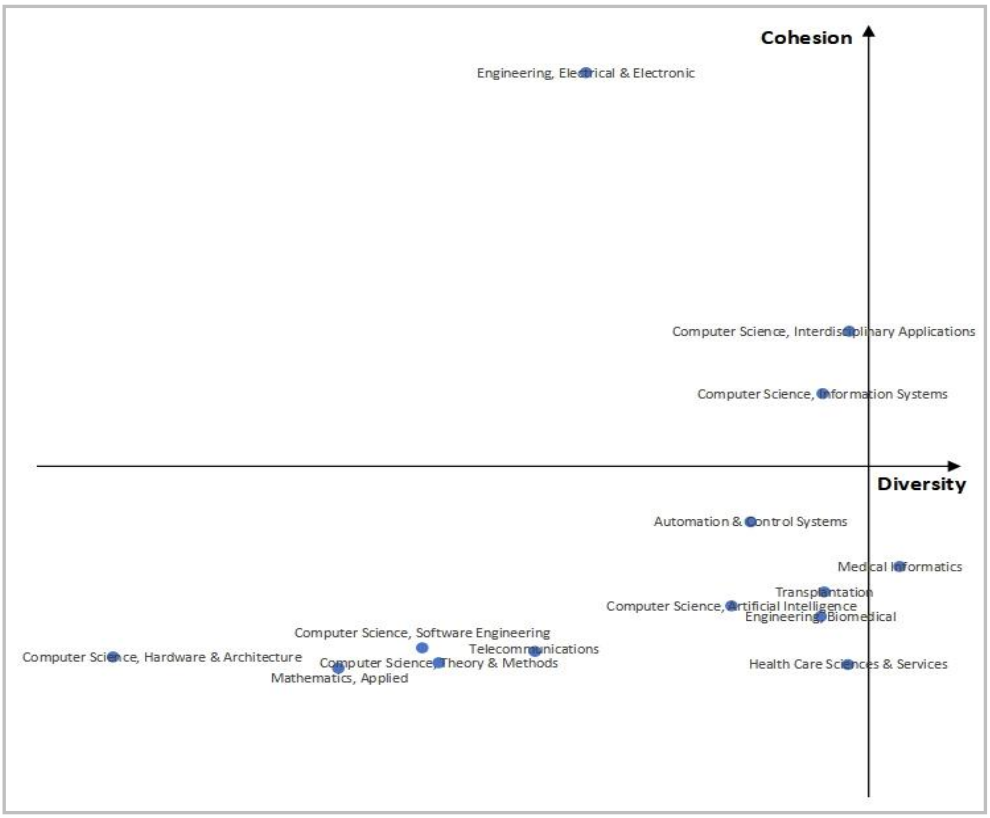


Figure 3. The two-dimensional feature analysis chart of the categories of publications (2004–2007).

Table 2. Two-dimensional Features of the categories of AI publications (2004–2007).

Two-Dimensional Characteristics	*Field of Categories*	*Category Names*
Quadrant I : High Cohesion - High Diversity (0 category)	None	None
Quadrant II : High Cohesion - Low Diversity (3 categories)	Computer Science and Engineering	Computer Science, Interdisciplinary Applications; Computer Science, Information Systems; Engineering, Electrical & Electronic
Quadrant III : Low Cohesion - Low Diversity (10 categories)	Computer Science and Engineering	Automation & Control Systems; Computer Science, Artificial Intelligence; Telecommunications; Computer Science, Theory & Methods; Computer Science, Software Engineering; Mathematics, Applied; Computer Science, Hardware & Architecture
	Health Care	Health Care Sciences & Services
	Medicine	Transplantation
	Chemistry & Applied Physics	Engineering, Biomedical
Quadrant IV : Low Cohesion - High Diversity (1 category)	Health Care	Medical Informatics
Categories with Betweenness Centrality of 0 (7 categories)	Computer Science and Engineering	Robotics; Engineering, Manufacturing; Transportation Science & Technology; Instruments & Instrumentation
	Pharmacology	Mathematical & Computational Biology
	Environmental Engineering	Energy & Fuels
	Chemistry & Applied Physics	Physics, Applied

From the two-dimensional feature analysis chart of the categories of publications, no categories were found in the high cohesion-high diversity quadrant (Quadrant I)

during this period. This indicates that Tongji University's interdisciplinary research in the field of artificial intelligence was still in infancy, and there was no widespread or close interconnection between disciplines at that time.

In the High cohesion-Low diversity quadrant (Quadrant II), three categories were distributed: "Computer Science, Interdisciplinary Applications," "Computer Science, Information Systems," and "Engineering, Electrical & Electronic," all of which belong to the field of Computer Science and Engineering. These categories had relatively high cohesion, indicating a strong ability for disciplinary integration. However, their diversity was low, meaning their knowledge absorption scope was relatively limited and primarily concentrated within closely related sub-disciplines of the same broader field. It is noteworthy that the field of Computer Science and Engineering exhibited strong disciplinary cohesion during this period, making it the most important bridge for communication between other disciplines and the formation of academic networks. In the context of the rapid digital transformation and technological advancements in information technology at the time, these fields had significant potential for interdisciplinary applications.

The low cohesion-low diversity quadrant (Quadrant III) included 10 categories, including "Computer Science, Artificial Intelligence," "Automation & Control Systems," "Health Care Sciences & Services," "Transplantation", etc. Spanning the fields of Computer Science and Engineering, Health Care, Medicine, Chemistry & Applied Physics. These categories were characterized by both low cohesion and low diversity, reflecting a narrower research citation range and insufficient interdisciplinary interaction. During this period, limited computational resources were a major bottleneck for the development of artificial intelligence. Compared to today's high-performance GPUs and distributed computing architectures, the computational capabilities of 2004-2007 were still limited, which restricted the training of complex models and the processing of large-scale data. As a result, the development of artificial intelligence was still in its early stages. The research focus of the "Computer Science, Artificial Intelligence" at this time was likely more concentrated on theoretical deepening within the discipline itself, rather than fostering interdisciplinary communication and applications. For instance, the most highly cited paper in the "Computer Science, Artificial Intelligence" field during this period was an article aimed to solve the of lack trust in P2P (Peer-to-Peer) Semantic Web (Wang, Zeng & Yuan, 2006). Despite not being directly related to artificial intelligence technologies or theories, solving this issue in P2P semantic networks not only enhances data reliability but also contributes to the development of distributed AI, knowledge graphs, and autonomous intelligent systems, thus providing a solid foundation for the innovation and popularization of artificial intelligence technologies.

In the low cohesion-high diversity quadrant (Quadrant IV), only the category of "Medical Informatics". "Medical Informatics" is the field that uses computers and related information technologies to handle tasks such as the storage, organization, retrieval, and optimal utilization of biomedical data, information, and knowledge, with the goal of supporting research and practice in the medical field and improving the accuracy, timeliness, and reliability of problem-solving and decision-making.

Between 2004 and 2007, the research in the "Medical Informatics" at Tongji University consisted of only two papers on the same topic, both published in 2005 (Xu et al.,2005; Xu et al.,2005). These studies used specific algorithms to train Artificial Neural Networks (ANNs) to identify the best treatment strategies for vascular tissue engineering based on experimental data. The results demonstrated the huge potential of artificial intelligence technology in decision-making within tissue engineering, enabling the analysis of large datasets and making more precise decisions to improve outcomes. These two papers cited 21 disciplines, spanning Computer Science and Engineering, Environmental Science, Biomedicine, Medicine, Biology, Agriculture and Food, Infectious Diseases, and Health Care. This wide span of disciplines explains the high measure of diversity for "Medical Informatics." However, due to the categories connected to "Medical Informatics" are themselves tightly interrelated, meaning these categories can communicate directly without needing "Medical Informatics" as an intermediary, resulting in a low measure of cohesion for "Medical Informatics".

statistically analysed the categories with betweenness centrality of 0, including "Robotics," "Mathematical & Computational Biology," "Energy & Fuels," and "Physics, Applied," indicating that they did not appear on the shortest connection path between any two categories. However, this does not entirely negate their potential intermediary value, as the flow of information between disciplines can also occur via non-shortest paths, and nodes along such paths can still facilitate knowledge exchange.

In summary, during 2004-2007, the level of interdisciplinarity in the fields of AI at Tongji University was weak. "Computer Science, Artificial Intelligence" during this period did not exhibit high diversity or cohesion. The number of categories in the Quadrant II and Quadrant III was high. The absence of categories in the Quadrant I suggests that the interdisciplinary nature of the AI research field had not yet reached a highly converged level. However, as academic collaboration deepens and disciplines development, categories located in the Quadrant II and IV may gradually evolve toward higher cohesion and diversity, thus driving the further development and evolution of the entire interdisciplinary research field in artificial intelligence.

The two-dimensional feature of AI research field during 2008-2011

The two-dimensional feature distribution of the categories in Tongji University's AI research field from 2008 to 2011 is presented in Figure 4 and Table 3. Compared to the situation from 2004 to 2007, the diversity and cohesion distribution of categories underwent significant changes, with the number of categories of publications increasing from 21 to 35. Among them, "Computer Science, Artificial Intelligence" achieved significant breakthroughs in diversity and cohesion, with the influence of AI in interdisciplinary research expanding continuously, gradually becoming a core driving force in the academic network.

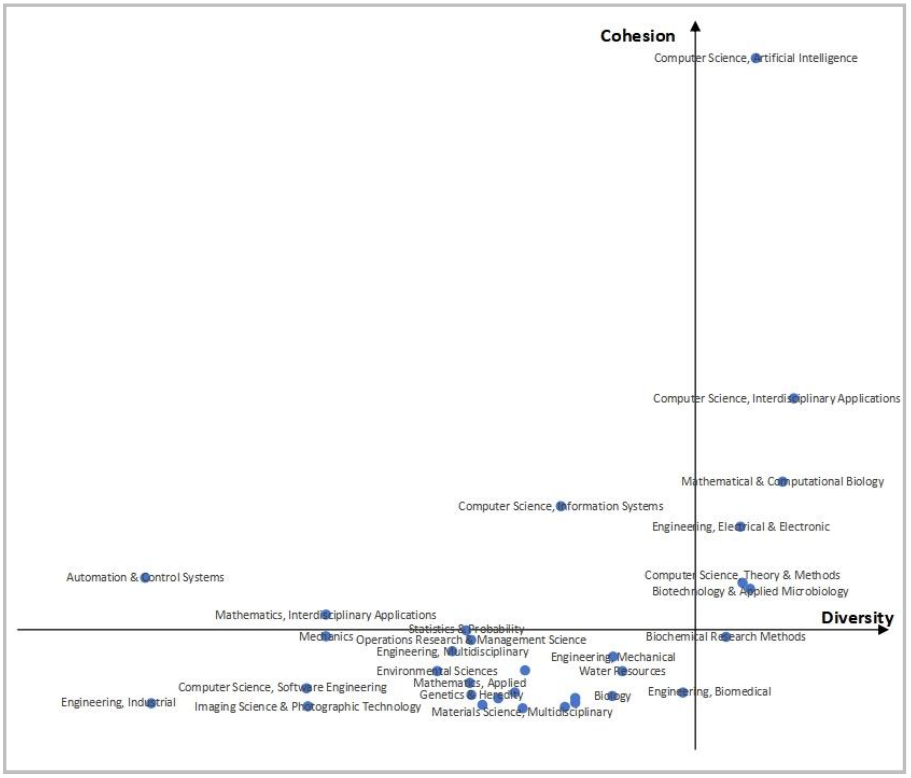


Figure 4. The two-dimensional feature analysis chart of the categories of publications (2008–2011).

Table 3. Two-dimensional Features of the categories of publications (2008–2011).

Two-Dimensional Characteristics	*Field of Categories*	*Category Names*
Quadrant I : High Cohesion - High Diversity (6 categories)	Computer Science and Engineering	Computer Science, Artificial Intelligence; Computer Science, Interdisciplinary Applications; Computer Science, Theory & Methods; Engineering, Electrical & Electronic
	Pharmacology	Mathematical & Computational Biology
	Agriculture and Food	Biotechnology & Applied Microbiology
Quadrant II : High Cohesion - Low Diversity (3 categories)	Computer Science and Engineering	Computer Science, Information Systems; Mathematics, Interdisciplinary Applications; Automation & Control Systems

Quadrant III : Low Cohesion - Low Diversity (22 categories)	Computer Science and Engineering	Computer Science, Hardware & Architecture; Telecommunications; Operations Research & Management Science; Mathematics, Applied; Statistics & Probability; Engineering, Multidisciplinary; Computer Science, Software Engineering; Engineering, Industrial
	Environmental Sciences	Water Resources; Remote Sensing; Environmental Sciences; Imaging Science & Photographic Technology
	Engineering	Engineering, Mechanical; Thermodynamics; Mechanics
	Chemistry & Applied Physics	Engineering, Biomedical; Materials Science, Multidisciplinary
	Chemistry	Chemistry, Multidisciplinary
	Biology	Biology
	Physics	Optics
	Pharmacology	Genetics & Heredity
	Medicine & Others	Medicine, Research & Experimental
Quadrant IV : Low Cohesion - High Diversity (1 category)	Chemistry	Biochemical Research Methods
Categories with Betweenness Centrality of 0 (3 categories)	Computer Science and Engineering	Robotics
	Engineering	Engineering, Aerospace
	Chemistry & Applied Physics	Physics, Applied

From the vector distribution of citing categories, the interdisciplinarity and cohesion levels of the artificial intelligence interdisciplinary field from 2008 to 2011 improved compared to 2004 to 2007. The number of categories in the Quadrant I increased from a previous vacancy to 6. These categories span across the fields of Computer Science and Engineering, Pharmacology, Agriculture and Food. Among them, "Computer Science, Artificial Intelligence" exhibited a significant increase in cohesion and diversity, making it the category with the highest cohesion level among all the categories in this phase. This indicates that the influence of the AI field has

substantially expanded, gradually establishing itself as a core hub in the academic network. "Computer Science, Interdisciplinary Applications" and "Engineering, Electrical & Electronic" made a significant leap from the Quadrant II (High Cohesion -Low Diversity) in the previous phase to the Quadrant I (High Cohesion -High Diversity). "Computer Science, Theory & Methods," which was in the Quadrant III (low cohesion-low diversity) from 2004 to 2007, and "Mathematical & Computational Biology," which had betweenness centrality of 0, both made a remarkable leap to the Quadrant I from 2008 and 2011, entering the realm of high cohesion−high diversity disciplines. "Biotechnology & Applied Microbiology" made its debut directly in the Quadrant I, emerging as a new force in the interdisciplinary field of artificial intelligence.

"Mathematics and Computational Biology" is a discipline that uses statistical methods and computer algorithms to analyze genetic and genomic data. It emphasizes research in molecular evolution, molecular classification, molecular genetics, and population genetics using mathematical, statistical, and computational approaches based on biological data. Compared to traditional biological experiments, which are limited by the precision of operational levels, experimental tools, and observational accuracy, computational biology based on computers is faster, more cost-effective, and theoretically has unlimited computational precision and high reproducibility. This characteristic not only significantly enhances the efficiency of biological research but also drives the interdisciplinary integration of bioinformatics, genomics, computer science, operations research, and other fields, exhibiting strong cross-disciplinary integration and accelerating the collaborative development and innovation in related fields (Mao, Jiang and Yuan, 2024). One of the most-cited papers in the 2008-2011 period in the field of artificial intelligence and "Mathematical & Computational Biology" at Tongji University discussed the application of decision tree methods in machine learning to biology, such as cancer classification and genomics classification (Che et al., 2011).

"Biotechnology & Applied Microbiology" is a comprehensive discipline that covers both "Biotechnology" and "Applied Microbiology," with "Applied Microbiology" being an important component of the "Biotechnology" discipline system. "Biotechnology" includes many biological programs adjusted according to human needs, such as early animal domestication, plant cultivation, and the improvement of varieties through artificial selection and hybridization. Under the modern scientific paradigm, biotechnology has evolved into gene engineering, cell culture, tissue culture, and other technologies. At the same time, many pure life sciences fields, such as biochemistry, cell biology, embryology, microbiology, and molecular biology, are also related to biotechnology. One of the most-cited papers in the 2008-2011 period in Tongji University's artificial intelligence interdisciplinary field of "Biotechnology and Applied Microbiology" designed a novel transcriptome assembly algorithm called IsoLasso based on RNA-Seq. This algorithm can simultaneously reconstruct all full-length mRNA transcripts from millions of short-read sequencing data, achieving higher sensitivity and accuracy than the most advanced transcriptome assembly tools. The study found that although this research did not directly involve artificial intelligence-related technologies and methods, the

breakthroughs made by the IsoLasso algorithm in short-read transcriptome assembly laid the groundwork for the later application of AI in gene sequencing (Li, Feng and Jiang, 2011).

In the Quadrant II (High Cohesion -Low Diversity), 3 categories were distributed in 2008-2011, including "Computer Science, Information Systems," "Mathematics, Interdisciplinary Applications," and "Automation & Control Systems".

In the Quadrant III (Low Cohesion -Low Diversity), the distribution of categories remains the most concentrated, with a total of 22 categories, an increase from the 2004-2007 period. This quadrant includes categories from nine fields: Computer Science and Engineering, Environmental Sciences, Engineering, Chemistry & Applied Physics, Chemistry, Biology, Physics, Pharmacology, Medicine & Others. This distribution suggests that although some disciplines have started to participate in interdisciplinary artificial intelligence research, their ability to integrate knowledge and their inclusivity remain at a relatively low level. They have yet to play a significant role in the cross-disciplinary knowledge flow network.

In the Quadrant IV (Low cohesion -High diversity), only "Biochemical Research Methods," was appeared during the 2008-2011 period. It is noteworthy that "Medical Informatics," which was in this quadrant in the previous phase, does not appear in any quadrant in this stage, indicating a significant disruption in the continuity of research in this field at Tongji University.

Furthermore, statistical results show that three categories with betweenness centrality of 0 during 2008-2011, including the newly added "Engineering, Aerospace" and the previously existing "Robotics" and "Physics, Applied." These disciplines are at the periphery of the academic network, contributing almost nothing to the overall network's connectivity and information flow. This phenomenon may be due to their low co-citation frequency with other disciplines, which has prevented the formation of effective knowledge transfer pathways. Additionally, related interdisciplinary research is still in its early stages and has yet to establish a stable network of academic interaction.

In summary, compared to the 2004-2007 period, the diversity and cohesion in the interdisciplinary field of artificial intelligence at Tongji University progress in 2008-2011. The Quadrant I show a breakthrough from non-existence to presence, with "Mathematics and Computational Biology" and ""Biotechnology & Applied Microbiology" becoming significant interdisciplinary forces outside the field of Computer Science and Engineering, playing a central role in knowledge integration within the AI cross-disciplinary field. However, most disciplines remain concentrated in the low cohesion -low diversity region (Quadrant III), with a relatively high number of peripheral disciplines, indicating that the level of communication between disciplines in the overall academic network has not yet fully matured. As interdisciplinary collaboration deepens, more disciplines are expected to transition from the Quadrant III to the Quadrant II or the Quadrant IV, and ultimately move toward the Quadrant I (high cohesion -high diversity), thus further advancing research in the AI interdisciplinary field.

The two-dimensional feature of AI research field during 2012-2015

The two-dimensional feature distribution of categories for the period 2012-2015 is shown in Figure 5 and Table 4. Compared to 2008-2011, the number of categories increased by 7, and the two-dimensional distribution of categories exhibited new changes.

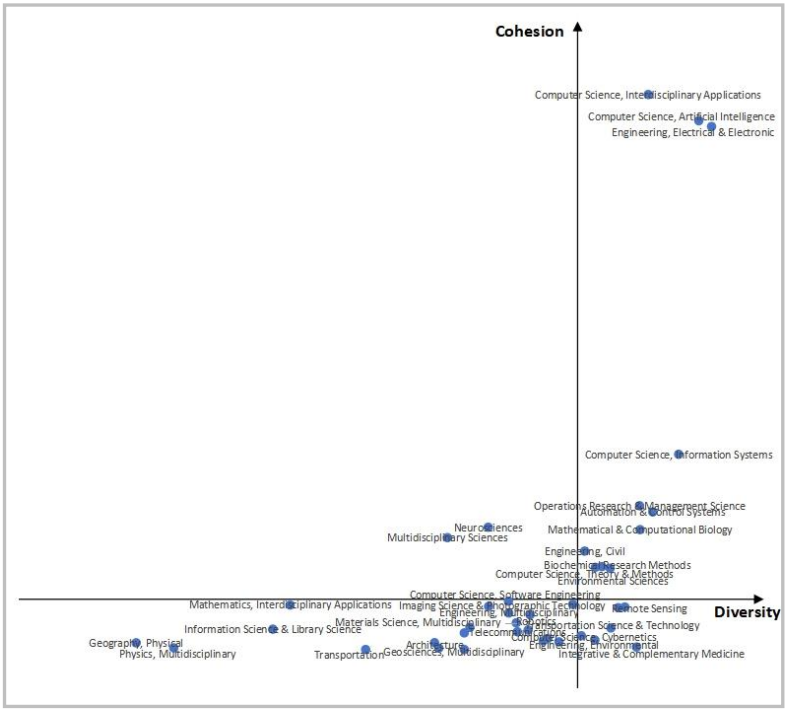


Figure 5. The two-dimensional feature analysis chart of the categories of publications (2012–2015).

Table 4. Two-dimensional Features of the categories of publications (2012–2015).

Two-Dimensional Characteristics	*Field of Categories*	*Category Names*
Quadrant I : High Cohesion -High Diversity (11 categories)	Computer Science and Engineering	Engineering, Electrical & Electronic; Computer Science, Artificial Intelligence; Computer Science, Information Systems; Automation & Control Systems; Computer Science, Interdisciplinary Applications; Operations Research & Management Science; Computer Science, Theory & Methods
	Engineering	Engineering, Civil
	Chemistry	Biochemical Research Methods

	Environmental Sciences	Environmental Sciences
	Pharmacology	Mathematical & Computational Biology
Quadrant II : High Cohesion - Low Diversity (2 categories)	Medicine & Others	Neurosciences; Multidisciplinary Sciences
Quadrant III : Low Cohesion - Low Diversity (21 categories)	Computer Science and Engineering	Robotics; Telecommunications; Engineering, Multidisciplinary; Computer Science, Software Engineering; Computer Science, Hardware & Architecture; Mathematics, Interdisciplinary Applications
	Environmental Sciences	Water Resources; Imaging Science & Photographic Technology; Geosciences, Multidisciplinary; Geography, Physical
	Engineering	Engineering, Mechanical; Construction & Building Technology; Architecture
	Physics	Optics; Physics, Multidisciplinary
	Management	Information Science & Library Science
	Chemistry	Chemistry, Multidisciplinary
	Chemistry & Applied Physics	Materials Science, Multidisciplinary
	Social Sciences	Transportation
	Biomedicine	Biochemistry & Molecular Biology
	Pharmacology	Pharmacology & Pharmacy
Quadrant IV : Low Cohesion - High Diversity (6 categories)	Computer Science and Engineering	Transportation Science & Technology; Computer Science, Cybernetics
	Pharmacology	Integrative & Complementary Medicine
	Health Care	Health Care Sciences & Services
	Environmental Sciences	Remote Sensing
	Environmental Engineering	Engineering, Environmental
Categories with Betweenness Centrality of 0 (2 categories)	Environmental Sciences	Geology
	Physics	Astronomy & Astrophysics

From 2012 to 2015, 11 categories entered the Quadrant I, including "Engineering, Electrical & Electronic," "Computer Science, Artificial Intelligence," "Engineering, Civil," "Biochemical Research Methods," "Environmental Sciences," "Mathematical & Computational Biology", etc. Compared to the 2008-2011 period, 6 additional categories were added. These categories span across five major fields: Computer Science and Engineering, Engineering, Chemistry, Environmental Sciences, Pharmacology.

We analyse some of the categories that entered the Quadrant I. The development of "Computer Science, Artificial Intelligence" is closely related to the technological breakthroughs in deep learning after 2012. Deep learning, as an emerging field of machine learning, aims to automatically extract multi-layer feature representations from data. Its core idea is to use a data-driven approach with a series of nonlinear transformations to extract features from raw data, progressing from low-level to high-level, from specific to abstract, and from general to specialized semantics (Zhang, Wang, & Guo, 2018). The landmark event in the development of deep learning was the 2012 ImageNet Large-Scale Visual Recognition Challenge (Russakovsky et al., 2015), where Krizhevsky's deep convolutional neural network model reduced image classification error rates by nearly 50% (MIT Technology Review, 2013). Compared to the best traditional methods in 2011, the recognition error rate dropped by 41.1%. By 2015, image recognition error rates based on deep learning had surpassed human performance. Since then, deep learning has been widely applied in fields such as speech recognition, image processing, and natural language processing, profoundly influencing research directions in the interdisciplinary field of AI. It has gradually evolved into a core node in the interdisciplinary network, with strong interdisciplinarity and disciplinary cohesion. These qualities not only drive technological breakthroughs in related fields but also strengthen AI's critical position in the academic network of interdisciplinary fields. From 2012 to 2015, there were 27 publications of "Computer Science, Artificial Intelligence" at Tongji University, citing a total of 70 categories across 17 major categories fields (out of a total of 19, with Environmental Engineering and Infectious Diseases being excluded). During this period, the most cited paper in the "Computer Science, Artificial Intelligence" focused on improving the robustness of image recognition through deep learning techniques. Specifically, the paper involved training an autoencoder to extract shape and color features of objects from RGB images. These features were then passed into a Recurrent Neural Network (RNN) to extract multi-level features, resulting in hierarchical and robust feature representations. Finally, the features from each subset were combined and sent to a SoftMax classifier for object recognition.

"Engineering, Electrical & Electronic" also performed notably during this period. Electricity is an excellent carrier of both energy and information; it can be used to collect, store, process, transmit, and present information, as well as distribute, store, and convert energy in various forms. In simple terms, electrical and electronic engineering is the modern method of managing information and energy. Devices, circuits, and systems form the three foundational concepts of electrical and electronic engineering. Devices are the basic elements that construct circuits, circuits serve as

carriers for specific functions, and systems are overarching architectures composed of multiple circuits to achieve complex objectives. This can be summarized as a transition from "hard" to "soft," where lower-level devices are closer to physics, while higher-level systems are more aligned with software and algorithms. As such, electrical and electronic engineering is a broad field that provides foundational technological support for numerous research and application areas. Examples include electronic systems widely used in aircraft and automobiles, precision instruments for medical diagnosis and surgery, wireless communication technologies enabling global connectivity, and semiconductor chip technologies supporting advancements in computing and AI. These applications illustrate the profound impact of electrical and electronic engineering on modern life and technological development.

"Operations Research & Management Science" also demonstrated high diversity and cohesion in the AI interdisciplinary field during 2012-2015, driven by its strong tool-based nature, broad application scenarios, and deep integration with AI and big data technologies. Operations research and management science focus on optimizing resource allocation and planning activities to maximize the utility of limited resources and achieve overall optimal objectives. To achieve this, mathematical methods are often employed to construct problem models, establish corresponding theories, and design and analyse solution algorithms. With advancements in computing and AI, computational capabilities have increased millions of times compared to traditional manual calculations, greatly expanding the application scenarios of operations research and management science. Its methods and theories are widely applied in fields such as engineering, economics, computer science, transportation, and supply chain management, significantly promoting interdisciplinary collaboration and knowledge integration.

Additionally, as a traditional strength of Tongji University, "Engineering, Civil" significantly enhanced its diversity and cohesion during this period by leveraging the rapid development of AI, a total of 4 papers was published, which collectively cited 25 different categories. Ji and Zhang (2012) utilized computer vision theory to establish the mathematical relationship between image planes and real-world space. This approach allowed for the capture of image sequences of planar targets mounted on vibrating structures, taken from digital cameras. By analyzing these images, the method could quantify structural dynamic displacements at the target positions, providing a precise measurement of complex object motion. This contributed significantly to structural displacement measurements in civil engineering projects. Computer vision is an important branch of artificial intelligence, focused on visual processing, which enables computers to "see" and understand visual content such as images and videos, mimicking human vision. The core capabilities of current computer vision systems are primarily based on deep learning models, which can process visual data from cameras, videos, or images. These capabilities include tasks such as image classification, object detection, pose estimation, image segmentation, and facial recognition. The deep integration of computer vision with machine learning has led to significant advancements in its applications across various fields.

"Mathematics & Computational Biology," as a typical representative of interdisciplinary research, continued to build on its advantages from the previous stage, further showcasing its critical role in the integration of bioinformatics and AI.
From the perspective of other quadrants, the number of categories in the Quadrant II decreased from 4 in 2008-2011 to 2 in 2012-2015. Among the 4 categories in the Quadrant II during 2008-2011, all except for "Mathematics, Interdisciplinary Applications" moved to the Quadrant I, while "Mathematics, Interdisciplinary Applications" shifted to the Quadrant III, reflecting a decline in its academic cohesion. In 2012-2015, the categories in the Quadrant II, "Neuroscience" and "Multidisciplinary Sciences," both from the Medicine field, had not appeared in the previous two stages but emerged prominently in this phase. These disciplines exhibit a relatively high level of academic cohesion, indicating their significant role in connecting knowledge within the artificial intelligence interdisciplinary field. However, due to their higher degree of specialization or the fact that their application scenarios have not yet become widely generalized, their interdisciplinary diversity remains relatively low.
In the Quadrant IV, there are 6 categories, including "Transportation Science & Technology," "Computer Science, Cybernetics," "Integrative & Complementary Medicine," "Health Care Sciences & Services," "Remote Sensing," and "Engineering, Environmental." These disciplines involve a broad range of interdisciplinary content in their research areas, but their internal cohesion remains weak, preventing them from forming tight structural connections within the academic network.
The Quadrant III remains the most populated quadrant, with 21 categories, similar to the number in 2008-2011. These categories include "Robotics," "Water Resources," "Engineering, Mechanical," "Optics", etc. Spanning across 10 major Categories Fields. Among these, "Robotics" published 6 papers in this phase, citing 24 different disciplines, but its academic influence remained concentrated in specific fields without deep collaboration with other disciplines in the AI interdisciplinary field. Consequently, its cohesion and diversity were relatively low.
In this phase, the disciplines "Astronomy & Astrophysics" and "Geology" with betweenness centrality of 0. However, it is important to note that these fields are increasingly influenced by AI, and their future potential for application and interdisciplinary collaboration is promising.
Overall, from 2012 to 2015, the cohesion and diversity of categories in the AI interdisciplinary field at Tongji University significantly increased. The number of disciplines in the Quadrant I expanded from 6 in the previous phase to 11. "Computer Science, Artificial Intelligence," promoted the interdisciplinary research and application of disciplines closely related to AI, such as "Engineering, Electrical & Electronic," "Operations Research & Management Science," "Engineering, Civil". Additionally, the number of disciplines in the second and fourth quadrants increased significantly, suggesting strong potential for future growth toward the first quadrant. Although most disciplines still cluster in the low cohesion−low diversity region (Quadrant III), this also provides new opportunities for academic development and innovation in the future.

The two-dimensional feature of AI research field during 2016-2019

The two-dimensional feature distribution of the categories in Tongji University's artificial intelligence interdisciplinary field from 2016 to 2019 is shown in Figure 6 and Table 5. During this period, artificial intelligence exhibited new characteristics, such as deep learning, cross-domain integration, human-machine collaboration, collective intelligence openness, and autonomous control, driven by new theories and technologies in mobile internet, big data, supercomputing, sensor networks, and brain science, as well as the strong demands of economic and social development. In 2017, China's State Council released the New Generation Artificial Intelligence Development Plan, explicitly emphasizing the need to seize major strategic opportunities in AI development. Subsequently, in 2018, the Ministry of Education issued the Artificial Intelligence Innovation Action Plan for Higher Education Institutions to further advance AI development. Against this policy backdrop, 2016-2019 became the initial phase of policy responses for AI development, and Tongji University's interdisciplinary AI research entered a new phase. The number of categories further increased to 102, and the AI interdisciplinary field gradually permeated more categories, showcasing a robust trend of interdisciplinary integration and development.

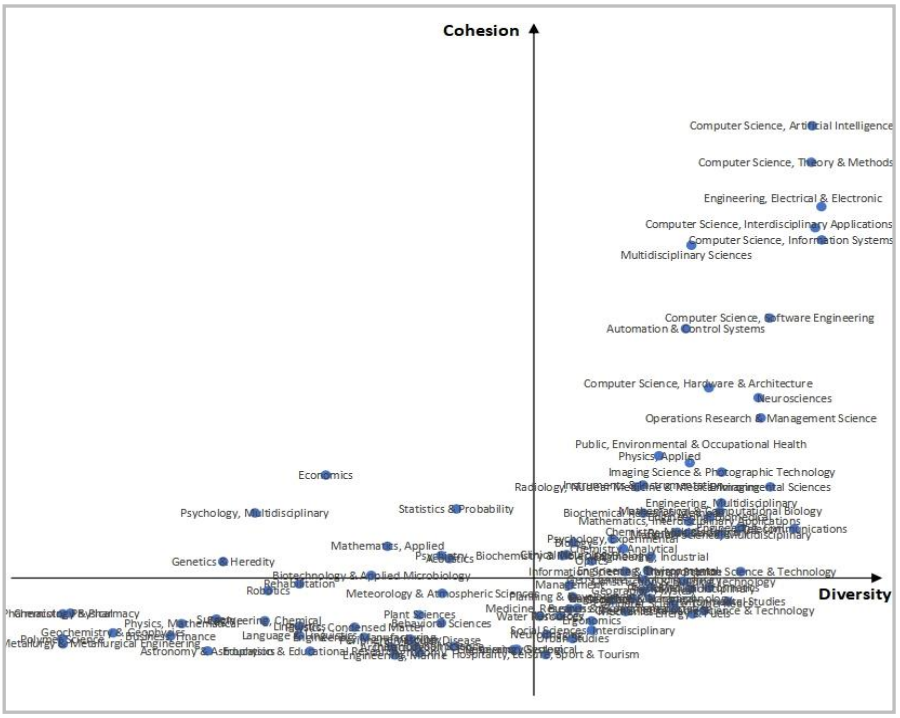


Figure 6. The two-dimensional feature analysis chart of the categories of publications (2016–2019). (To ensure clarity in the visualization given the high number of nodes, only the labels of key nodes are retained in the chart).

Table 5. Two-dimensional Features of the categories of publications (2016–2019).

Two-Dimensional Characteristics	*Field of Categories*	*Category Names*
Quadrant Ⅰ: High Cohesion - High Diversity (37 categories)	Medicine & Others	Neurosciences; Multidisciplinary Sciences
	Medicine	Radiology, Nuclear Medicine & Medical Imaging; Clinical Neurology
	Health Care	Public, Environmental & Occupational Health
	Pharmacology	Mathematical & Computational Biology
	Psychology	Psychology, Experimental
	Physics	Optics
	Biomedicine	Biochemistry & Molecular Biology
	Biology	Biology; Optics
	Computer Science and Engineering	Computer Science, Information Systems; Engineering, Electrical & Electronic; Computer Science, Interdisciplinary Applications; Computer Science, Artificial Intelligence; Computer Science, Theory & Methods; Telecommunications; Computer Science, Software Engineering; Operations Research & Management Science; Transportation Science & Technology; Engineering, Multidisciplinary; Computer Science, Hardware & Architecture; Mathematics, Interdisciplinary Applications; Automation & Control Systems; Engineering, Industrial; Instruments & Instrumentation
	Environmental Sciences	Environmental Sciences; Imaging Science & Photographic Technology; Remote Sensing

	Environmental Engineering	Engineering, Environmental
	Chemistry & Applied Physics	Materials Science, Multidisciplinary; Engineering, Biomedical; Physics, Applied
	Chemistry	Chemistry, Multidisciplinary; Biochemical Research Methods; Chemistry, Analytical
	Management	Information Science & Library Science
	Engineering	Engineering, Civil
Quadrant II : High Cohesion - Low Diversity (8 categories)	Computer Science and Engineering	Statistics & Probability; Mathematics, Applied
	Engineering	Acoustics
	Psychology	Psychiatry; Psychology, Multidisciplinary
	Agriculture and Food	Biotechnology & Applied Microbiology
	Social Sciences	Economics
	Pharmacology	Genetics & Heredity
Quadrant III : Low Cohesion - Low Diversity (29 categories)	Environmental Sciences	Engineering, Geological; Soil Science; Meteorology & Atmospheric Sciences; Geochemistry & Geophysics
	Medicine	Respiratory System; Behavioral Sciences; Peripheral Vascular Disease; Surgery
	Agriculture and Food	Agronomy; Plant Sciences
	Engineering	Thermodynamics; Engineering, Marine; Architecture
	Biomedicine	Hematology
	Computer Science and Engineering	Engineering, Manufacturing; Robotics
	Chemistry & Applied Physics	Physics, Condensed Matter; Chemistry, Physical; Metallurgy & Metallurgical Engineering
	Psychology	Education & Educational Research; Rehabilitation; Language & Linguistics; Linguistics
	Environmental Engineering	Engineering, Chemical

	Physics	Astronomy & Astrophysics; Physics, Mathematical
	Social Sciences	Business, Finance
	Pharmacology	Pharmacology & Pharmacy
	Chemistry	Polymer Science
Quadrant IV：Low Cohesion - High Diversity (26categories)	Social Sciences	Environmental Studies; Transportation; Geography; Urban Studies
	Health Care	Medical Informatics
	Environmental Engineering	Green & Sustainable Science & Technology; Energy & Fuels
	Physics	Physics, Multidisciplinary
	Engineering	Construction & Building Technology; Engineering, Mechanical; Mechanics; Planning & Development
	Computer Science and Engineering	Computer Science, Cybernetics
	Biology	Ecology
	Chemistry & Applied Physics	Nanoscience & Nanotechnology
	Environmental Sciences	Geosciences, Multidisciplinary; Geography, Physical; Water Resources
	Management	Ergonomics; Social Sciences, Interdisciplinary; Management; Business; Hospitality, Leisure, Sport & Tourism
	Medicine & Others	Medicine, Research & Experimental
	Biomedicine	Oncology
	Medicine	Neuroimaging
Categories with Betweenness Centrality of 0 (2 categories)	Computer Science and Engineering	Logic
	Social Sciences	Social Issues

From the perspective of the distribution in a two-dimensional quadrant, from 2016 to 2020, the number of categories located in the Quadrant I increased significantly from 11 to 37 compared to the previous phase, includes 15 Fields of Categories. It is the first time, the Quadrant I became the one with the most categories, breaking the longstanding dominance of the Quadrant III in terms of category numbers.

During this phase, the influence of "Neuroscience" in the interdisciplinary field of artificial intelligence at Tongji University significantly increased. Neuroscience and artificial intelligence are closely related fields. In 1945, John von Neumann, in a paper outlining the architecture of modern digital computers, proposed that "the operation of the nervous system is", in fact, "digitally encoded on the surface," thus suggesting that the brain could inspire the development of computers. For example, in artificial intelligence, some principles used in Artificial Neural Networks (ANNs) are inspired by neuroscience. These include Convolutional Neural Networks (corresponding to the visual cortex), Regularization (corresponding to steady-state plasticity), Max Pooling (corresponding to lateral inhibition), Dropout (corresponding to synaptic failure), and Reinforcement Learning, which reflect the synergistic interaction between artificial intelligence and neuroscience. From 2016 to 2019, "Neuroscience" published a total of 7 papers at Tongji University, which cited 61 categories across 15 categories fields. The most cited paper during this period was an article published in 2018, it based on brain CT data from patients and healthy individuals, designed a novel 14-layer Convolutional Neural Network (CNN) for the early detection and identification of multiple science (MS), achieving an overall accuracy of 98.23% in the detection results (Wang et al., 2018).

The number of categories in the high cohesion-low diversity quadrant (Quadrant II) increased from 2 in the previous phase to 8, including "Statistics & Probability," "Applied Mathematics," "Acoustics," "Psychiatry," "Psychology, Multidisciplinary," "Biotechnology & Applied Microbiology," "Economics," and "Genetics & Heredity." These disciplines did not appear in the previous phase but directly entered the high cohesion category in this phase, indicating that once these disciplines emerged, they played a crucial role in connecting the academic network. However, their disciplinary diversity in terms of citation range remains relatively weak.

The number of categories in the low cohesion-high diversity quadrant (Quadrant IV) also increased significantly, from 1 in the previous phase to 26. These include "Environmental Studies," "Medical Informatics," "Green & Sustainable Science & Technology," "Physics, Multidisciplinary," etc., covering 13 categories fields. Although these categories are not prominent in terms of their intermediary position in the academic network, they are characterized by a large number and wide span of cited disciplines in their research, making them representative of interdisciplinary integration in the field of AI.

The number of disciplines in the Quadrant III (low cohesion -low diversity) ranks second to the Quadrant I, with a total of 29 disciplines, an increase of 8 from the previous phase. These disciplines are mostly from traditional fields or applied research areas with more limited scope. They include Biomedicine fields such as "Respiratory System," "Haematology," "Pharmacology & Pharmacy," Science application field such as "Geotechnical Engineering," "Thermodynamics," "Manufacturing Engineering," "Chemical Engineering," as well as foundational fields such as physics, chemistry, and mathematics, and social sciences such as "Education & Educational Research," "Business, Finance," and "Linguistics." Many of these are new categories that were not addressed in previous phases, with weaker

interdisciplinary integration and cohesion, resulting in limited synergistic effects in the cross-disciplinary field of artificial intelligence.

In this phase, only "Logic" and "Social Issues" had 0 betweenness centrality, and both are considered "new categories" in this study. These categories were included precisely because they have gradually been influenced by artificial intelligence-related fields. It is foreseeable that, with the further expansion and application of AI technologies, these disciplines may strengthen their interdisciplinary collaborations with other fields in the future, showcasing greater potential for integration.

Overall, from 2016 to 2019, driven by artificial intelligence-related policies in China, Tongji University's interdisciplinary field of artificial intelligence also saw new developments. More exploratory interdisciplinary research was conducted, such as the integration of artificial intelligence and foundational fields like neuroscience. During this phase, the overall diversity and cohesion of disciplines in the artificial intelligence field at Tongji University were further enhanced. The number of disciplines in the Quadrant I surpassed that in the Quadrant III for the first time, and the number of categories in the Quadrant IV was much greater than in the Quadrant II. This suggests that high cohesion disciplines usually have higher disciplinary diversity in their citations, but disciplines with high diversity in cited fields may not necessarily have strong connectivity within the citation network. The core interdisciplinary development of artificial intelligence disciplines also made significant progress, not only broadening the scope of disciplines but also enhancing the overall connectivity of the academic network, thereby laying a solid foundation for deeper interdisciplinary collaboration and application in the future.

The two-dimensional feature of AI research field during 2020-2023

Between 2020 and 2023, the interdisciplinary field of artificial intelligence (AI) experienced rapid growth under the combined influence of policy support, technological advancements, and social demand. The number of categories increased sharply, rising from 102 categories in the previous phase to 172 categories, with a more diverse range of categories. This reflects the significant impact of AI technology in driving development across related fields.

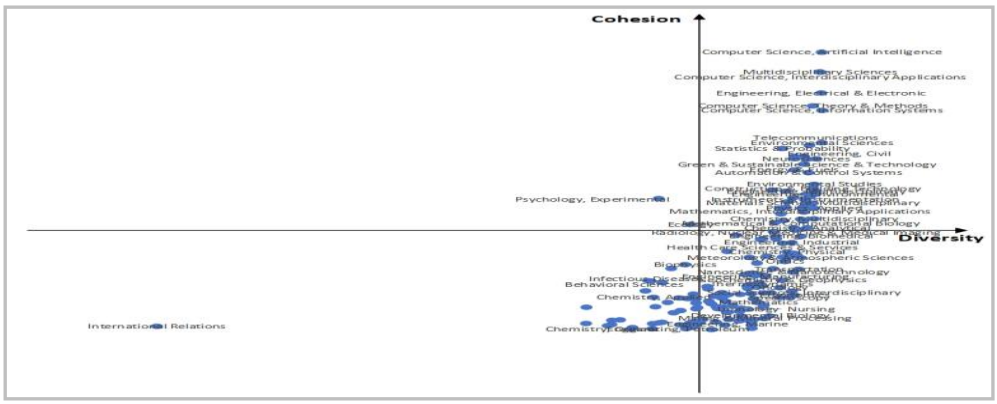


Figure 7. The two-dimensional feature analysis chart of the categories of publications (2020–2023).

First, the number of categories in the high cohesion-high diversity quadrant (Quadrant I) continued its growth, reaching 59 categories. These covered fields such as Computer Science and Engineering, Environmental Sciences, medicine, Engineering, and social sciences, encompassing nearly all of Tongji University's key development areas. During this phase, AI technologies demonstrated stronger disciplinary integration capabilities. Particularly with breakthroughs in deep learning, automation and control systems, and AI applications in the medical field, the ability of AI to drive interdisciplinary convergence significantly increased.

Second, compared to the previous phase, the overall distribution of disciplines in terms of the two-dimensional vector shows a clear shift to the right, indicating that the disciplines are accelerating their development towards higher diversity. At the same time, the number of disciplines in the Quadrant IV was significantly higher than in other quadrants, with 86 disciplines, whereas the Quadrant II (high cohesion−low diversity) only had 2 disciplines. This further validates the asymmetry of the two-dimensional development of cohesion and diversity characteristics, namely that categories with high cohesion tend to have high diversity levels, while high diversity disciplines do not necessarily exhibit high cohesion.

Notably, it is worth noting that during this phase, no disciplines with betweenness centrality of 0. This indicates that all disciplines in the literature have, to varying degrees, connected with other disciplines within the interdisciplinary network of artificial intelligence, further emphasizing the overall improvement in the connectivity of Tongji University's AI cross-disciplinary field and the deepening of interdisciplinary collaboration.

In conclusion, from 2020 to 2023, driven by relevant policies, Tongji University’s interdisciplinary field of artificial intelligence entered a period of explosive development. During this phase, the number of disciplines increased significantly, and the overall disciplinary distribution shifted towards high diversity and high cohesion. The degree of interdisciplinary integration further deepened. From the two-dimensional distribution of disciplines in the literature, it is evident that artificial intelligence has had a broad and profound impact on multiple disciplinary fields. It not only holds a central position in computer science but has also deeply infiltrated fields such as computational biology, medicine, civil engineering, urban planning, transportation, environmental science, and agriculture in the natural sciences and engineering technologies. Additionally, AI has influenced social sciences, including psychology, education, sociology, and international relations, forming a pattern driven by artificial intelligence at its core, characterized by deep interdisciplinary integration. This underscores the key role of artificial intelligence in promoting disciplinary convergence and innovative development.

Conclusion

We used Tongji University's publications on artificial intelligence as a case to study AI's interdisciplinarity with the indicators of the diversity and cohesion of the references of these publications. We collected the AI publications of Tongji University and sliced them into five four-year periods, then calculated every period's

indicators of the categories' diversity and cohesion, and plotted these categories in a Cartesian coordinate system.

On the whole, from 2004 to 2023, the number of categories in the AI research field has significantly increased. The overall distribution of categories has shifted from low cohesion and low-diversity to high-cohesion and high-diversity. This phenomenon indicates that AI has had a broad and profound impact on multiple academic fields, and is deeply penetrated in the natural sciences and engineering disciplines such as computational biology, medicine, civil engineering, urban planning, transportation, environmental science, and agriculture. Additionally, AI has influenced social science fields including psychology, education, sociology, and international relations. This has led to a pattern where AI serves as the core driving force for the deep interdisciplinary convergence and integration of multiple disciplines, highlighting its critical role in promoting interdisciplinary convergence and innovative development.

From the perspective of key nodes, "Computer Science, Artificial Intelligence" initially resided in the Quadrant III (low cohesion-low diversity), directly transitioned to the Quadrant I (high cohesion-high diversity) during 2008-2011, where it has since maintained its high cohesion and high diversity characteristics. In addition, categories such as "Engineering, Electrical & Electronic," "Mathematics & Computational Biology," "Biotechnology & Applied Microbiology," "Operations Research & Management Science" and "Engineering, Civil" had already entered the high cohesion-high diversity category before 2016, demonstrating strong interdisciplinary integration capabilities. Examining the development history of these categories reveals the critical role AI technologies have played in driving their growth. "Neuroscience" and artificial intelligence are closely related fields., showed a significant increase in its academic influence after 2016, reflecting the synergistic interaction between artificial intelligence and neuroscience.

The research also found that the development of cohesion and diversity levels of disciplines exhibits asymmetry. Specifically, categories with high cohesion tend to have higher diversity, and are more likely to enter the high cohesion -high diversity quadrant. However, categories with higher diversity do not necessarily exhibit strong cohesion.

This study also counts the categories with a betweenness centrality of 0. Although these categories do not serve as "bridges" or "mediators" in connecting any two other categories, and contribute almost nothing to the overall network connectivity and information flow, the trend in their numbers indicates that they have gradually transitioned from isolation to integration in the field of AI. Furthermore, they have increasingly been influenced by artificial intelligence, suggesting significant potential for interdisciplinary applications and collaboration in the future.

Overall, this study demonstrates that artificial intelligence is not only a highly interdisciplinary field but also a comprehensive and leading catalytic force that can deeply merge with and permeate various disciplines. This power could be applied and validated within other disciplines. Therefore, in the development of disciplines at universities, it is essential to align with the trend of AI-enabled interdisciplinary integration and actively explore the fusion of AI with traditional disciplines. For

instance, civil engineering is evolving toward smart construction, mechanical engineering is advancing toward intelligent manufacturing, and transportation is progressing toward smart mobility. AI research should transcend disciplinary boundaries, opening new journeys within the intersection and fusion of disciplines, and moving toward broader frontiers.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.72274139).

References

Bai, J., Wu, Y., Zhang, J., & Chen, F. (2015). Subset based deep learning for RGB-D object recognition. Neurocomputing, 165, 280-292.

Cai, S., Wang, Q., & Shen, Y. (2020). Artificial intelligence empowerment: Innovation and development of university discipline construction-An interview with Professor Chen Jie, Academician of the Chinese Academy of Engineering. E-Learning Research, 41(02), 5–9.

Che, D., Liu, Q., Rasheed, K., & Tao, X. (2011). Decision tree and ensemble learning algorithms with their applications in bioinformatics. Software tools and algorithms for biological systems, 191-199.

Chubin, D. E., Porter, A. L., & Rossini, F. A. (1984). "'Citation classics'" analysis: An approach to characterizing interdisciplinary research. Journal of the American Society for Information Science, 35(6), 360–368.

Diego, C., Puay, T., & Ismael, R. (2014). Interdisciplinarity and research on local issues: Evidence from a developing country. Research Evaluation, 23(3), 195–209.

Dogan, M., & Pahre, R. (1990). Creative marginality: Innovation at the intersections of social sciences. Westview Press.

Ji, Y. F., & Zhang, Q. W. (2012, July). A novel image-based approach for structural displacement measurement. In Proc. 6th Int. Conf. Bridge Maintenance, Safety Manage (pp. 407-414).

Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. Journal of the American Society for Information Science and Technology, 58(9), 1303–1319.

Liang, D., Bo, W. H., & Jiang, L. B. (2017). Computational biology: An emerging frontier discipline in biology. Chinese Forestry Education, 35(S1), 139–142.

Liu, Y. (2018). Constructing global backbone of science based on inter-categories co-membership of journals. Journal of the China Society for Scientific and Technical Information. 37(06): 580-589.

Liu, Y., Rafols, I., & Rousseau, R. (2012). A framework for knowledge integration and diffusion. Journal of Documentation, 68(1), 31-44.

Liu, Y., Rousseau, R. (2010). Knowledge diffusion through publications and citations: a case study using ESI-Fields as unit of diffusion. Journal of the American Society for Information Science and Technology. 61(2): 340–351.

Li, W., Feng, J., & Jiang, T. (2011). IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. Journal of computational biology: a journal of computational molecular cell biology, 18(11), 1693–1707.

Mao, K. Y., Jiang, Y., Yuan, Y. C., et al.. (2024). Development trends of computational biology in 2023. Life Sciences, 36(1), 11–20.

MIT Technology Review. (2013, April 23). The 10 breakthrough technologies of 2013 [EB/OL]. Retrieved from https://www.technologyreview.com/s/513981/the-10-breakthrough-technologies-of-2013/

Pan, L. B., Ge, L. Q., & Wang, S. (2022). The development of information systems: From network-driven to knowledge-driven. Journal of China Academy of Electronics and Information Technology, 17(9), 929–934.

Porter, A. (2006). Interdisciplinary research: Meaning, metrics and nurture. Research Evaluation, 15, 187–195.

Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. Scientometrics, 82, 263–287.

Rafols, I., Leydesdorff, L., O'"Hare, A., et al. (2012). How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. Research Policy, 41(7), 1262–1282.

Rousseau, R., Zhang, L., & Hu, X. (2019). Knowledge integration: Its meaning and measurement. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), Springer Handbook of Science and Technology Indicators (pp. 69-94). Springer Verlag.

Russakovsky, O., Deng, J., Su, H., et al. (2015). ImageNet large-scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211–252.

Stirling, A. (2007). A general framework for analysing diversity in science, technology and society. Journal of the Royal Society Interface, 4, 707–719.

Von Neumann, J. (1993). First draft of a report on the EDVAC. IEEE Annals of the History of Computing, 15(4), 27-75.

Wang, S. H., Tang, C., Sun, J., Yang, J., Huang, C., Phillips, P., & Zhang, Y. D. (2018). Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling. Frontiers in neuroscience, 12, 818.

Wang, W., Zeng, G., & Yuan, L. (2006, September). A semantic reputation mechanism in P2P semantic web. In Asian Semantic Web Conference (pp. 682-688). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wu, Z. H. (2019). Promoting talent training and technological innovation in artificial intelligence through convergence. China University Teaching, No. 342(2), 4–8.

Xu, J., Ge, H., Zhou, X., Yan, J., Chi, Q., & Zhang, Z. (2005). Prediction of vascular tissue engineering results with artificial neural networks. Journal of Biomedical Informatics, 38(6), 417-421.

Xu, J., Zhou, X., Yang, D., Ge, H., Wang, Q., Tu, K., & Guo, T. (2005). Applying informatics in tissue engineering. Methods of Information in Medicine, 44(01), 38-43.

Zhang, J. Y., Wang, H. L., & Guo, Y. (2018). A review of deep learning research. Computer Applications Research, 35(7), 1921–1928+1936.

Zhang, L., Sun, B. B., & Huang, Y. (2020). Interdisciplinary scientific research: Connotation, measurement, and impact. Science Research Management, 41(7), 279–288.

https://doi.org/10.51408/issi2025_029

Interdisciplinarity, Collaboration and Industry Links in Australian Discovery and Linkage Projects (2023-25)

Hamid R. Jamali

h.jamali@gmail.com
School of Information and Communication Studies, Charles Sturt University, Wagga Wagga, NSW (Australia)

Abstract

Research funding schemes play an important role in shaping national research priorities and facilitating collaboration between academia and industry. This study examines collaboration patterns and interdisciplinarity in Australian Research Council (ARC) Discovery and Linkage projects funded between 2023 and 2025. Using network analysis and visualisation techniques, the study analyses data from 1,750 projects (315 Linkage and 1,435 Discovery) to investigate interstate collaboration patterns, disciplinary interactions, and industry engagement in Australian research. Interdisciplinary Distance (IDD) values were calculated using Field of Research (FoR) codes, incorporating both disciplinary disparity and balance of disciplines involved in each project to compare interdisciplinarity of the two project types. The analysis reveals distinct patterns in interstate collaboration, with eastern states demonstrating stronger collaborative ties. Engineering and Biological Sciences dominate both grant types. Linkage projects show greater interdisciplinarity than Discovery projects, with 53.7% of Linkage projects involving three distinct FoR codes compared to 48.1% of Discovery projects. IDD values of Linkage projects (mean = 0.46) were also significantly higher than Discovery projects (mean = 0.38). Analysis of industry partners in Linkage projects 2024 reveals concentration in Public Administration and Safety (21.1%) and Professional Services (19.5%) sectors, primarily in New South Wales and Victoria. The findings highlight opportunities and suggest policy implications for geographical and sectoral diversification in research collaboration, particularly in engaging underrepresented regions and industries.

Introduction

The research landscape in Australia is rich with a strong higher education sector (42 universities) that is its economy's fourth largest export sector (Spre, 2023), and high-quality research with more than 90 per cent of its university research rated as world class or higher (Universities Australia, 2019). However, there are also problems such as insufficient collaboration (Cetindamar et al., 2024), low rate of university-industry partnership (Jackson et al., 2018) and inability to translate research and innovation into innovative products (Jackson et al., 2016).

To address such challenges and direct research efforts, governments worldwide employ competitive funding schemes as key policy instruments. In Australia, this takes the form of a complex national competitive funding scheme that is managed by two research councils: The Australian Research Council (ARC) and The National Health and Medical Research Council (NHMRC). These councils operate various grant programs that each serves specific strategic objectives including supporting promising early career researchers and developing research infrastructure to fostering scientific discoveries and facilitating knowledge exchange with industry. The ARC's Discovery and Linkage schemes, in particular, have become crucial instruments for directing research efforts and stimulating collaboration, with some

disciplines such as humanities showing strong reliance on these schemes for research funding (Turner and Brass, 2014).

Understanding the dynamics of collaboration, interdisciplinarity, and industry engagement within these grants is useful for evaluating their broader impact on Australia's research ecosystem. Collaboration across states and territories can reveal geographical patterns of academic networking. The co-occurrence of FoR codes provides insights into the interdisciplinary (could be also considered as multi-disciplinary) nature of funded projects. Furthermore, the connections between industry sectors and academic disciplines highlight the role of Linkage grants in bridging the gap between research and practice.

Therefore, my aim in this study is to examine these interrelated aspects using data from ARC-funded projects for funding years 2023, 2024 and 2025. By analysing inter-state collaborations, interdisciplinary patterns, and industry partnerships, this research seeks to answer the following research questions.

- What are the patterns of inter-state collaboration within Australian Discovery and Linkage projects, and how do they differ?
- What is the extent and nature of interdisciplinarity in Discovery and Linkage projects, and how do these differ?
- What are the primary industry sectors engaged in Linkage projects, and how do these sectors connect to specific Fields of Research?

About Discovery and Linkage projects

ARC Linkage projects are meant to create an alliance between industry and universities mediated by the government. They are specifically aimed at facilitating collaboration between researchers and industry, government, and community organisations, and helping with knowledge exchange between these organisations. They emphasise practical applications of research and aim to address real-world challenges, drive innovation, and create economic, social, and environmental benefits. Therefore, a condition of a Linkage project is that they must have industry partners (industry is broadly defined here) that contribute to the funding of the project. In contrast, Discovery projects focus on fundamental research and encourage the generation of new knowledge and theoretical advancements across a wide range of disciplines. Collaboration, national or international, is encouraged in all ARC projects and might increase the chance of success for grant applications.

Literature review

There have been some studies in the past on ARC-funded grants and projects. While some have focused on analysing grants awarded in specific fields, such as social work (Tilbury et al., 2020), religious studies (Possamai et al., 2021) or accounting (Clarke et al., 2011), in terms of topics investigated or the success rate of getting grants, others have used qualitative methods such as interviews to investigate other aspects such as knowledge co-production (e.g., Cherney, 2015).

A very relevant study to this one is an older study by Maldonado and Brooks (2004) who used an economic model to find out if research-intensive organisations (i.e.,

companies with high R&D and IP assets) are more likely to participate in Linkage projects. They found that high R&D expenditure and revenue increased the likelihood of an organisation's participation, but the role of IP (e.g., patents, designs) was not significantly related to Linkage participation. They also looked at the rate of participation by industry sector and found that sectors where a high social impact is perceived had a greater tendency to collaborate in projects, while sectors with a trading focus had a lower tendency to collaborate. Sectors with high participation rates include libraries, museums, and the arts, community services, water supply, sewerage and drainage services, forestry and logging, oil, and gas extraction, and rail transport. Sectors with low participation include several sectors such as insurance, finance, basic material wholesale, general construction, textiles, clothing, and so on.
A few other studies have found some characteristics of different grant types and their investigators. Discovery grant recipients have significantly higher citation counts than Linkage grant recipients (Brooks and Byrne, 2006) which might be because Discovery grants emphasis academic research and discovery; and a better academic research performance is expected from investigators.
Interdisciplinarity of research has been extensively researched and there are different methods and approaches to defining it and measuring it, for instance based on journals, citations, topics and so on. Depending on these, interdisciplinarity could also be considered multi-disciplinarity or diversity. For instance, in this study I am using FoR codes, similar to Bromham et al. (2016) and if a project has multiple FoR codes, we can argue that it is both interdisciplinary and multidisciplinary. Jamali et al (2020) reviewed different ways of measuring interdisciplinarity or diversification of research. Yegros-Yegros et al. (2015) measured three dimensions of interdisciplinarity including variety, balance and disparity and found that while variety had a positive effect on citation impact, balance and disparity had a negative effect. A key study on interdisciplinarity is the study by Bromham et al. (2016) that looked at FoR codes of 18,476 ARC Discovery grant proposals, both successful and unsuccessful ones submitted over five years and found that interdisciplinary proposals had less chance of success. They used FoR codes assigned by researchers to proposals to calculate an interdisciplinary score for each proposal.
Geographical distance/proximity plays a role in collaboration and over the years there have been many studies in this area (e.g., Frenken, Hardeman, and Hoekman, 2009) that suggest an increase in collaboration (and in a way in the globalisation of research). However, many of such studies concern international collaboration whereas here my focus is collaboration within Australia. A significant study in this area is a large-scale study by Lin, Frey and Wu (2023) that analysed 20 million articles and 4 million grants and found that ‘across all fields, periods and team sizes, researchers in remote teams are consistently less likely to make breakthrough discoveries relative to their on-site counterparts’ (p. 987). Regarding collaboration with industry, an older study by Ponds et al (2007) analysed article co-authorship data and found that the collaboration between different kinds of organisations is more geographically localised than collaboration between organisations that are similar due to institutional proximity. A more recent study and its review of the literature

indicates that generally geographic proximity is an important factor in university-industry collaboration (Alpaydin and Fitjar, 2021).

Methods

I used publicly available data from the Australian Research Council (ARC) API to analyse patterns in Linkage and Discovery grants. I obtained the data in JSON format, converted them into an SQLite database, and managed it using DB Browser for SQLite.

For Linkage Projects, I classified 'Partner Organisations' according to the Australian and New Zealand Standard Industrial Classification (ANZSIC) 2006 version (Australian Bureau of Statistics, 2006-revision-2.0). I used the D&B Hoovers database (a business database that provides information on companies and industries) to query organisation names and obtain their respective ANZSIC classifications. ANZSIC is a hierarchical classification with four levels. The database provides classification at lower detailed levels (3^{rd} or 4^{th}). For instance, 'BHP Group Limited' has the class '0801 - Iron Ore Mining' which is a fourth level code (called Classes in the classification). As there are many classes, I aggregated the classes into higher levels and used the top level which are divisions. For organisations that were not present in the D&B Hoovers database, I manually classified them based on their documented activities and by comparing them with similar organisations in the same sector.

ARC funded projects use FoR codes based on the Australian and New Zealand Standard Research Classification (ANZSRC) (Australian Bureau of Statistics, 2020). Grants commencing from 2023 onwards have been classified using ANZSRC 2020, while grants from 2007 to 2022 used ANZSRC 2008. Therefore, I focused on grants with funding commencement years from 2023 to 2025. In grant applications, researchers can assign up to three FoR codes at the six-digit subdivision level to their grant, with one designated as the primary code. While the percentage allocation across these codes sums to 100% for each grant, these specific percentages are not publicly available. Hence, in this study, I used FoR codes without their weighting (percentage value). To facilitate more meaningful analysis, I aggregated the six-digit codes to their corresponding two-digit hierarchical levels.

Moreover, to better understand the connections between FoRs and industry sectors, I grouped FoR codes and industry sectors into broader, meaningful categories. FoR codes (see Table 2 for their list) were grouped as follows:

- Arts and Humanities: 36, 43, 47, 50
- Social Sciences: 33, 35, 38, 39, 44, 45, 48, 52
- Life Sciences: 30, 31, 41
- Medical Sciences: 32, 42
- Physical Sciences: 34, 37, 40, 46, 49, 51

I grouped industry divisions (see Table 4 for their list) based on typical economic models, which classify industries by their role in the production and delivery of goods and services. Primary industries, for instance, are involved in resource

extraction, second industries are involved in manufacturing and infrastructure and so on.

- Primary Industries: A, B
- Secondary Industries: C, D, E
- Tertiary Industries (Services): F, G, H, I, J, K, L, M, N, O
- Social and Community Services: P, Q, R
- Other: S, NA

To assess the interdisciplinarity of the projects, I employed the Interdisciplinary Distance (IDD) metric based on the methodology developed by Bromham et al. (2016). The IDD quantifies both the disparity and balance of disciplines involved in a project by utilising a hierarchical classification of research fields similar to phylogenetic trees in evolutionary biology. I adapted this approach by applying a simplified correlation matrix. The simplified correlation included these values: Different Domain: 0.1 correlation; Same Domain, different Division: 0.3 correlation; Same Division, different Group: 0.6 correlation; Same Group, different Field: 0.8 correlation; Same Field: 1.0 correlation. This metrics allows to effectively measure and compare the degree of interdisciplinarity across ARC Discovery and Linkage projects. For a comprehensive description of the IDD calculation and its application, please see Bromham et al. (2016), which details the use of phylogenetic species evenness to standardise IDD scores between 0 (single-disciplinary) and 1 (maximum disparity with even representation). Further specifics on IDD can be found in the Supplementary Information of Bromham et al. (2016).

I used Python scripts to transform the grant data into network files suitable for visualisation and analysis. These network files were used to examine various relationships, including interstate collaborations, the co-occurrence of FoR codes across projects, and links between various industry sectors and FoR codes. I used ChatGPT to assist with writing the Python codes; however, I checked the accuracy of the codes and their outputs. For visualisation purposes, I used VOSviewer and SankeyMATIC.

Findings

I analysed the data of 1,435 Discovery and 315 Linkage projects, a total 1,750 grants awarded for funding commencement years 2023 to 2025. It is important to note that while there is only one round of applications for Discovery grants annually, there are two rounds of applications each year for Linkage grants. The number of 2025 Linkage grants is lower (56) because the outcome of the second round was not yet available at the time of data collection (December 2024). Table 1 shows the number of grants by type and year with some of their characteristics.

Table 1. Number of ARC-funded projects per year, and average number of investigator and organisation per project.

Year	*Project type*	*Number of grants*	*Ave number of investigators*	*Ave number of organisations*
2023	Discovery	478	3.47	2.36
2023	Linkage	137	5.61	4.38
2024	Discovery	421	3.39	2.24
2024	Linkage	122	5.83	4.47
2025	Discovery	536	3.40	2.18
2025	Linkage	56	6.25	4.55

When examining the average number of investigators and organisations involved in each project type, it becomes evident that Linkage projects exhibit higher participation in both aspects. This is unsurprising, as Linkage projects necessitate the inclusion of industry partners, who are designated as Partner Organisations within the grant applications. Furthermore, these partnerships often involve investigators from the partner organisations, who are then classified as Partner Investigators within the project.

Distribution of projects by Field of Research

Table 2 shows the distribution of projects by Field of Research divisions (two-digit FoR codes) for Discovery and Linkage projects from 2023 to 2025. Linkage projects for 2024 are also presented separately in a column, as industry analysis was done only on those projects.

Table 2. Number of projects (2023-2025) by Field of Research divisions based on primary FoR codes.

Field of Research Divisions (2-digit FoR)	*Discovery 2023-25*		*Linkage 2023-25*		*Linkage 2024 only*		*All*
	N	%	N	%	N	%	N
30 - Agricultural, Veterinary and Food Sciences	10	0.7	16	5.1	5	4.1	26
31 - Biological Sciences	259	18.0	20	6.3	8	6.6	279
32 - Biomedical and Clinical Sciences	54	3.8	4	1.3	2	1.6	58
33 - Built Environment and Design	18	1.3	7	2.2	2	1.6	25
34 - Chemical Sciences	87	6.1	18	5.7	9	7.4	105
35 - Commerce, Management, Tourism and Services	31	2.2	7	2.2	4	3.3	38
36 - Creative Arts and Writing	6	0.4	3	1.0	1	0.8	9
37 - Earth Sciences	44	3.1	10	3.2	3	2.5	54
38 - Economics	29	2.0	4	1.3	1	0.8	33
39 - Education	23	1.6	7	2.2	4	3.3	30
40 - Engineering	297	20.7	105	33.3	39	32.0	402

41 - Environmental Sciences	34	2.4	21	6.7	14	11.5	55
42 - Health Sciences	10	0.7	7	2.2	0	0	17
43 - History, Heritage and Archaeology	37	2.6	6	1.9	2	1.6	43
44 - Human Society	89	6.2	15	4.8	7	5.7	104
45 - Indigenous Studies	11	0.8	4	1.3	2	1.6	15
46 - Information and Computing Sciences	83	5.8	29	9.2	9	7.4	112
47 - Language, Communication and Culture	42	2.9	11	3.5	1	0.8	53
48 - Law and Legal Studies	20	1.4	3	1.0	2	1.6	23
49 - Mathematical Sciences	76	5.3	1	0.3	0	0	77
50 - Philosophy and Religious Studies	17	1.2	0	0	0	0	17
51 - Physical Sciences	87	6.1	5	1.6	3	2.5	92
52 - Psychology	71	4.9	12	3.8	4	3.3	83
Total	1,435	100	315	100	122	100	1,750

The distribution of grants across different FoRs is uneven. Engineering (FoR 40) and Biological Sciences (FoR 31) received disproportionately larger numbers of Discovery grants. Although Biological Sciences received more Linkage grants than many other fields, their number (N = 20) is far fewer than that of Engineering (N = 105), which received almost three times the number of grants as the second-largest Linkage receiver, FoR 46 (Information and Computer Sciences) (N = 29).
In 2024, there were 122 Linkage projects, with the largest number (N = 39) going to Engineering (FoR 40), followed by Environmental Sciences (N = 14). All FoR divisions received some Discovery grants, including Creative Arts and Writing, which had the smallest number (N = 6). However, this is not the case for Linkage projects, as Philosophy and Religious Studies (FoR 50) received no Linkage grants at all. Mathematical Sciences (FoR 49) had only one Linkage project.

Inter-state collaboration

Australia is a vast country, and its population (about 26 million) and universities (42 universities) are not evenly distributed geographically or across its eight states and territories. The east coast of Australia, which includes the three major states of New South Wales, Victoria, and Queensland, hosts most of its population and universities. By examining interstate collaboration, we can identify how different states and territories are connected and whether there are any notable differences between Linkage and Discovery projects. It should be noted that the names of states for each participating organisation are provided in ARC data only for Australian institutions in Discovery projects. For Linkage projects, such data is not provided for industry partners, even if they are Australian. Therefore, the first two visualisations below only include collaborations between Australian institutions (usually higher education institutions) and exclude industry partners for Linkage projects.

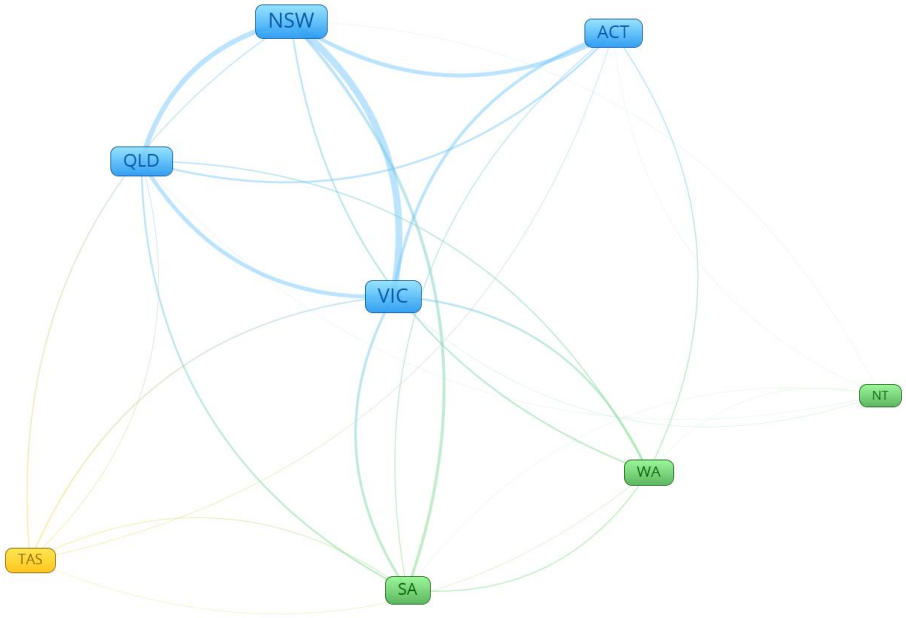


Figure 1. Inter-states collaborations in Discovery projects 2023 to 2025.

The network of interstate collaborations in Discovery projects (2023–2025) reveals three distinct clusters of research partnerships across Australia. The first cluster (Figure 1) comprises the eastern states (New South Wales, Victoria, Queensland) and the Australian Capital Territory (ACT), with NSW showing the strongest total link strength (736), followed by Victoria (644) and Queensland (430). These are major states with large metropolitan areas and several large Australian universities known as the Group of Eight. The second cluster includes Western Australia, South Australia, and the Northern Territory, with lower link strengths (164, 248, and 14, respectively). These states receive fewer grants. Northern Territory and South Australia are neighbouring states of Western Australia and this might influence their higher rate of collaboration. Tasmania forms a third cluster with the lowest connection weight (86). It has weak connections to all other states except for the Northern Territory (no link). Overall, it seems there is a trend of east coast universities mostly collaborating with one another, while universities on the west, north, and south coasts mostly collaborate within their regions.

Figure 2 shows the network of collaborations in Linkage projects (2023–2025). It depicts a more integrated collaboration pattern than Discovery projects, with seven out of eight states and territories forming a single cluster. Again, all states and territories have at least one link with one another, except that there is no link between Tasmania and the Northern Territory (similar to Discovery). New South Wales demonstrates the strongest collaborative intensity, with a total link strength of 370, followed closely by Victoria (328) and Queensland (264). The strongest link is between Victoria and NSW, with 143 connections. NT forms its own cluster due to its limited participation in cross-state collaboration and its smaller number of projects.

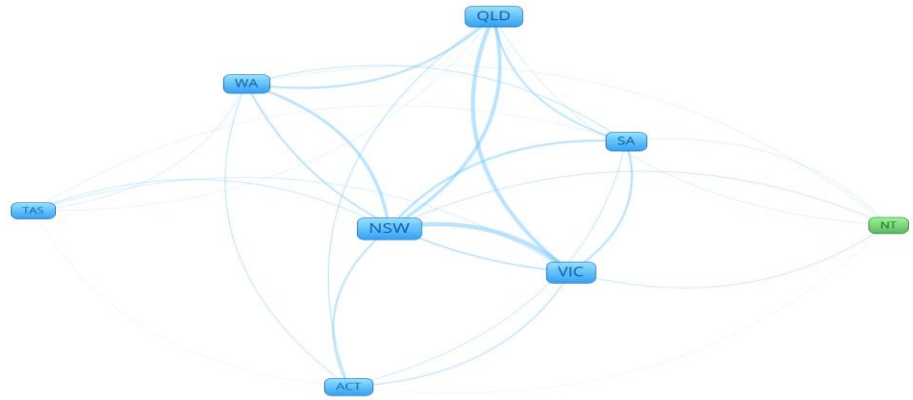


Figure 2. Inter-states collaborations in Linkage projects 2023 to 2025 (excluding industry partners).

For Linkage projects in 2024, state data were obtained from the D&B Hoovers database. Figure 3 shows the interstate collaboration only between Administering Organisations (the institution of the chief investigator) and Partner Organisations (industry partners). Most of partner organisations were concentrated in NSW (32.4%) and Victoria (18.9%). Queensland (13.2%), WA (12.8%) and SA (12.5%) had smaller number of partner organisations. ACT had 5.7% and the other two, i.e., NT and TAS each hosted less than 3% of industry partners. The network reveals two distinct clusters, with varying levels of industry engagement across states and territories. NSW demonstrates the strongest industry partnerships, with a total link strength of 238, and its strongest link is with Victoria (36). All eight states and territories have at least one link, except for Tasmania and the ACT, which have no links. South Australia and the ACT form a separate cluster, with link strengths of 96 and 74, respectively. The Northern Territory and Tasmania, despite being part of the main cluster, show the lowest intensity of interstate links (40 and 34, respectively).

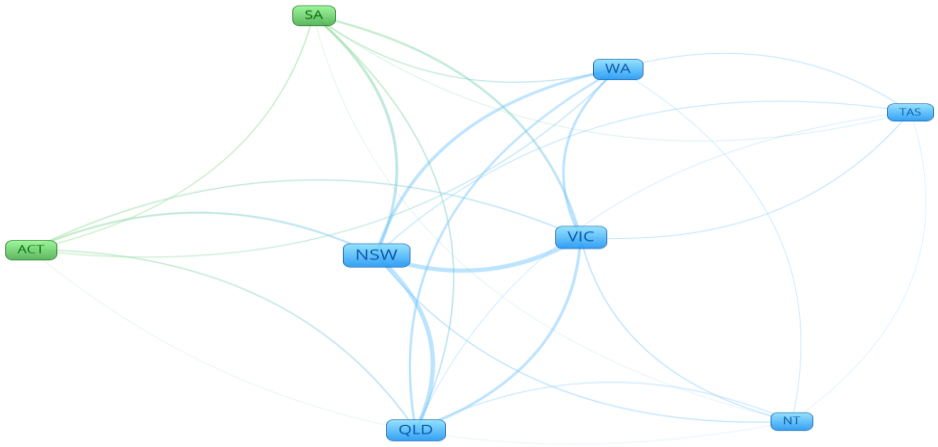


Figure 3. Inter-state collaboration between Administering Organisation and Industry Partners in Linkage 2024.

Interdisciplinarity of projects

The FoR codes (up to three) that investigators can assign to their projects at the time of grant application indicate how interdisciplinary (or multi-disciplinary) their project is and which Fields of Research are involved. Table 3 shows the number and percentage of projects in each project type that had one, two, or three distinct FoR codes at the six-digit, four-digit, and two-digit levels. It is clear that investigators of Linkage projects are slightly more likely to assign multiple FoR codes to their projects compared to Discovery projects. While 18.3% of Discovery projects had only one FoR code (six-digit), and 48.1% had three different FoR codes, in Linkage projects, 15.9% had only one FoR code, and 53.7% had three FoR codes. When aggregating the codes to higher levels of the FoR hierarchy, Linkage projects were still more likely to cover more FoR codes, with 9.8% of them having three distinct two-digit FoR codes, while this number was about half (4.7%) for Discovery projects.

The interdisciplinary distance (IDD) values shown in Figure 4 also confirm the findings above. IDD values are larger for Linkage projects which indicate Linkage projects are consistently more interdisciplinary. A Mann-Whitney U test revealed a statistically significant difference in IDD values between Discovery projects and Linkage projects ($U = 189882.0$, $Z = -4.496$, $p < 0.001$). The 95% confidence intervals for the mean of IDD were lower (0.363, 0.391) for Discovery projects compared to Linkage projects (0.425, 0.491).

Table 3. Number and % of projects with 1, 2 and 3 distinct 6-, 4- and 2-digit FoR codes.

	6-digit FoR			*4-digit FoR*			*2-digit FoR*		
Type	FoR	N	%	FoR	N	%	FoR	N	%
Discovery	1	263	18.3	1	548	38.2	1	977	68.1
	2	482	33.6	2	566	39.4	2	390	27.2
	3	690	48.1	3	321	22.4	3	68	4.7
Linkage	1	50	15.9	1	89	28.3	1	177	56.2
	2	96	30.5	2	125	39.7	2	107	34.0
	3	169	53.7	3	101	32.1	3	31	9.8

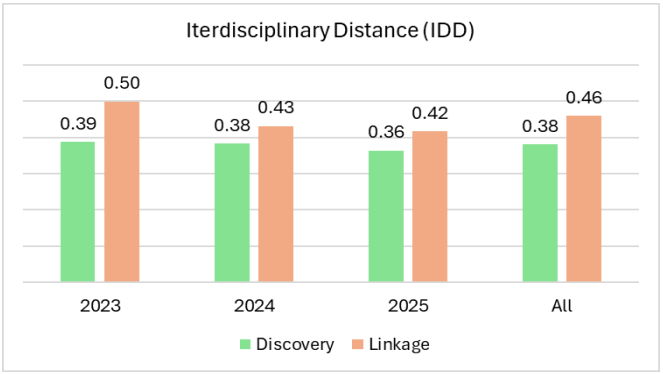


Figure 4. Interdisciplinary Distance (IDD) of Discovery and Linkage projects.

Looking at the network structure, in the case of Discovery projects (Figure 5), four clusters appear, showing clear interdisciplinary patterns. Cluster 1 in blue (on the left), with the largest number of FoR codes (13 of them), comprises mostly fields that can be considered as social sciences (e.g., 33, 35, 38) and arts and humanities (36, 43, 47, and 50), except for 42 (Health Sciences). The second cluster, in green, with five FoR codes, represents all physical sciences. Earth Sciences (FoR 37), however, forms its own cluster in the middle (Cluster 4). Cluster 3, in amber, with four FoR codes, is mostly life sciences (30, 31, and 41). FoR 32 (Biomedical and Clinical Sciences) can be considered part of Medical Sciences (together with Health Sciences 42). However, it is part of Cluster 3, which is dominated by life sciences. Although the clusters align with broad disciplinary groupings, the four clusters are well-connected. For instance, Information and Computer Sciences (46), which is part of physical sciences in Cluster 2, also has strong links with social sciences and humanities in Cluster 1, as well as life sciences in Cluster 3. This may be due to the wide application of computer science across various fields. The top three strongest links between FoRs in this network are between 31 and 32 (i.e., biological sciences, and biomedical and clinical sciences with 59 links), 34 and 40 (i.e., chemical sciences and engineering with 47 links), and 31 and 34 (i.e., biological, and chemical sciences with 36 links).

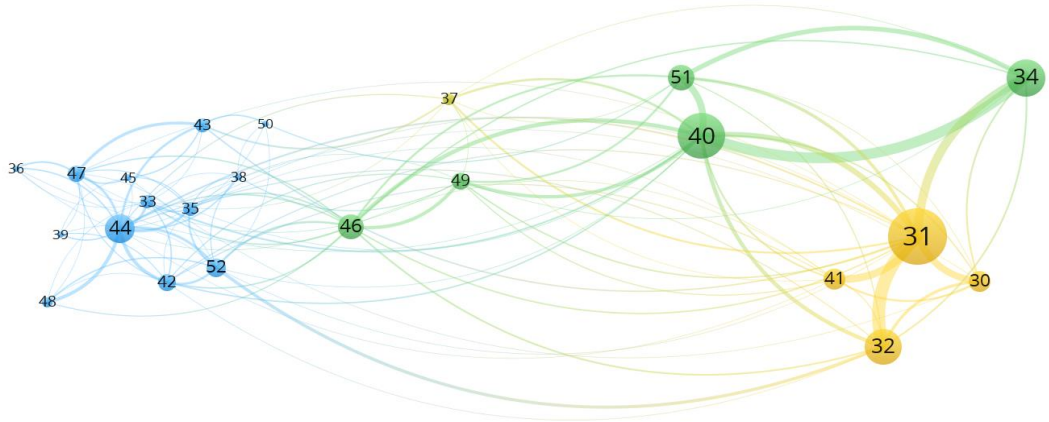


Figure 5. Co-occurrence of 2-digit FoR codes in Discovery projects 2023 to 2025.

The interdisciplinarity network for Linkage projects, illustrated in Figure 6, differs from that of Discovery projects. First, unlike Discovery projects that cover all 23 FoR codes, the Linkage network includes only 22 of them. FoR code 50 – Philosophy and Religious Studies – is absent from the network because no Linkage projects have used this code. Second, the Linkage network consists of five clusters, and these clusters do not follow the broad subject groupings seen in the Discovery network.
Cluster 1, in blue on the left, includes eight FoR codes, three of which are arts and humanities fields (36, 43, 47) that appear close to one another on the far left. The remaining five are social sciences (39, 44, 45, 48, 52). Cluster 2, in green at the top, contains five codes and is a mix of social sciences (33, 38) and physical sciences (40, 46, 51). Cluster 3, in amber, consists of four codes and combines life sciences (30, 31), medical sciences (32), and physical sciences (34). Cluster 4, in sulphur yellow, includes three codes: 35 from social sciences, 42 from medical sciences, and 49 from physical sciences. Finally, Cluster 5, in purple, includes two codes: 37 (earth sciences) from physical sciences and 41 (environmental sciences) from life sciences. Although the clusters are interconnected, the pattern shows that the co-usage of FoR codes in projects involving industry collaboration is different from that in projects focused primarily on scientific discovery. The three strongest links between FoRs in this network are between 31 and 41 (i.e., biological, and environmental sciences with 16 links), 40 and 41 (i.e., engineering and environmental sciences with 14 links), and 34 and 40 (i.e., chemical sciences and engineering with 10 links).

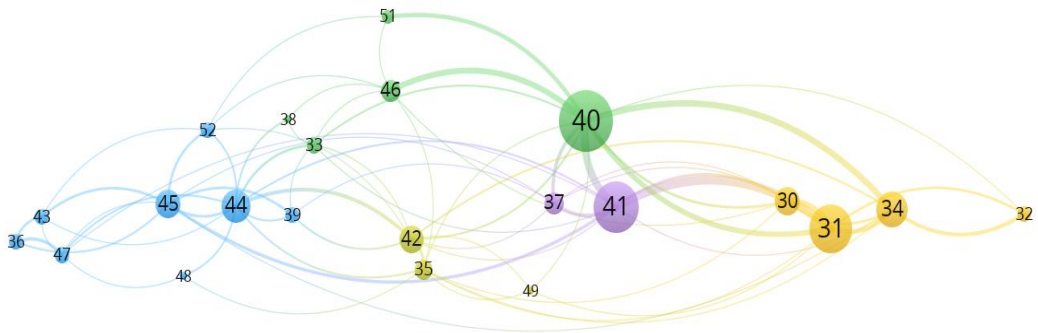


Figure 6. Co-occurrence of 2-digit FoR codes in Linkage projects 2023 to 2025.

Industry sectors

The analysis of partner organisations by industry sector, presented in Table 4, reveals that Public Administration and Safety (67 partners) and Professional, Scientific and Technical Services (62 partners) are the dominant sectors engaging in Linkage projects. Health Care and Social Assistance (37 partners) and Manufacturing (31 partners) also show strong representation. These are the industries that benefit more from Australian higher education as universities seem to focus more on solving the challenges of these sectors. Some sectors, such as Accommodation and Food Services and Rental, Hiring and Real Estate Services, show no participation.

Table 4. Number of Partner Organisation in Linkage projects 2024 from each industry sector.

ANZSIC Industry Divisions	*N*	*%*
A - Agriculture, Forestry and Fishing	5	1.6
B - Mining	11	3.5
C - Manufacturing	31	9.7
D - Electricity, Gas, Water and Waste Services	8	2.5
E - Construction	5	1.6
F - Wholesale Trade	4	1.3
G - Retail Trade	2	0.6
H - Accommodation and Food Services	0	0.0
I - Transport, Postal and Warehousing	1	0.3
J - Information Media and Telecommunications	3	0.9
K - Financial and Insurance Services	6	1.9
L - Rental, Hiring and Real Estate Services	0	0.0
M - Professional, Scientific and Technical Services	62	19.5
N - Administrative and Support Services	4	1.3
O - Public Administration and Safety	67	21.1
P - Education and Training	20	6.3
Q - Health Care and Social Assistance	37	11.6
R - Arts and Recreation Services	10	3.1
S - Other Services	29	9.1
NA - Non-classifiable Establishments	13	4.1
Total	318	100

The Sankey (alluvial) diagram (Figure 7) illustrates the connections between primary 2-digit FoR codes and industry sectors in 2024 Linkage projects. The diagram depicts 105 flows between 38 nodes, with a total of 318 links. The numbers next to FoR codes represent the sum of links or partner organisations associated with each code, while the numbers next to industry codes indicate the number of partner organisations in each industry.

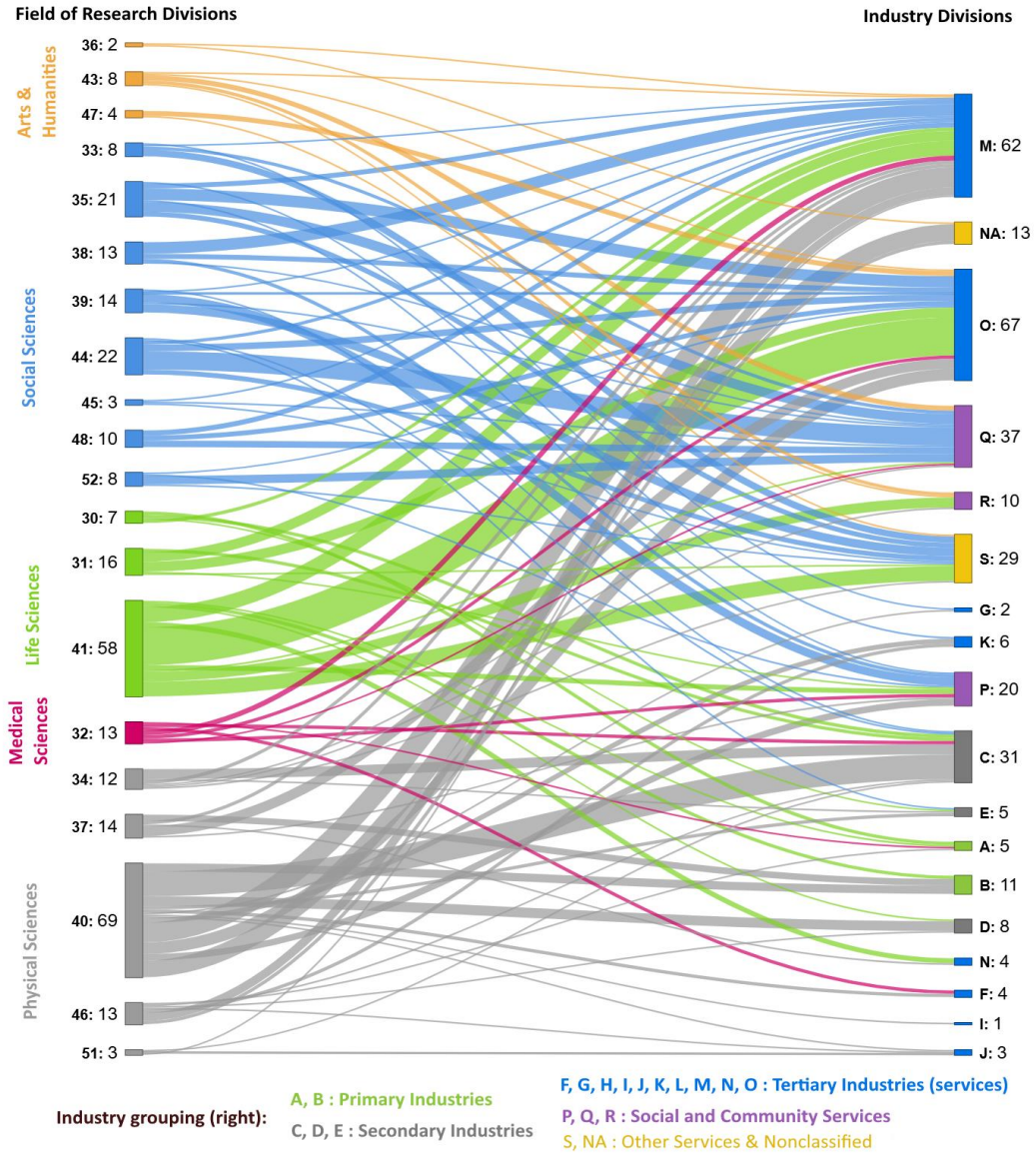


Figure 7. Sankey diagram of links between various primary FoR codes and industry sectors in Linkage projects.

The diagram reveals the patterns of research-industry engagement. Physical sciences, including engineering, demonstrate the broadest range of industry connections, linking with multiple sectors such as Manufacturing, Public Administration and Safety, and Professional Services. Life sciences, including environmental sciences, show strong connections with tertiary industries (services), including public sector organisations.

The Sankey diagram illustrates the connections between Primary FoR and industry sectors in Linkage projects. A key finding is the dominance of specific industries, such as tertiary industries (e.g., services categorised under M and O), which exhibit the highest level of engagement across multiple research fields. For instance, FoR

40 (Physical Sciences) is strongly connected to tertiary industries with 15 links to industry C and 12 links to M. Similarly, life sciences, particularly FoR 41, show substantial connections, prominently linking to industry O (23 links) and tertiary services more broadly.

The diagram also reveals the distribution of less prominent but meaningful connections, which indicate a degree of interdisciplinarity. Fields like arts and humanities (e.g., FoR 36) have limited but focused collaborations, such as their ties to industry M (professional … services). The medical sciences (e.g., FoR 32) display more diverse but smaller-scale collaborations across primary, secondary, and tertiary industries. Additionally, there are some niche yet significant links, such as FoR 40's (Engineering) connection to secondary industries like C (manufacturing) and D (electricity…).

Discussion

The analysis of ARC Discovery and Linkage projects reveal a few patterns in Australia's research funding landscape, particularly regarding collaboration patterns, disciplinary focus, and industry engagement. These patterns have implications for research policy and practice in Australia.

Industry partners are not distributed evenly across the states with NSW hosting by far more than other states, followed by Victoria. This to some extent should be expected as these are larger states in terms of population and there are probably more business and organisations located in them. However, there might be also room for wider inter-state collaboration with industry partners. In terms of industry sector, the largest was Public Administration and Safety (with 21.1%) followed by Professional, Scientific and Technical Services (with 19.5%), and Q - Health Care and Social Assistance (with 11.6%). All other sectors had less than 10% presence. Public Administration and Safety (class O) is mostly government agencies including local government (e.g. councils). The distribution of sectors shows some room for more diversity.

The diversity of participating organisations has increased over time. While in early years of the Linkage scheme, there were a small number of firms that actively involved in Linkage projects (Maldonado and Brooks, 2004), in 2024 there were many different organisations (about 300) involved in linkage projects from different sectors, although some sectors had a bit of dominance. This might indicate a positive change in Linkage projects and possibly increased awareness among potential industry partners about the benefits of academic collaboration. This is something that requires further investigation.

Both Discovery and Linkage grants are considered a type of Category One grant in Australia, that is to say they are the most prestigious and competitive and sought after by researchers. But Linkage applicants might not need to have a strong citation and publication track record as Discovery applicants have (Brooks and Byrne, 2006). This distinction reflects the different objectives of these grant schemes. Linkage projects emphasise practical impact and industry engagement over traditional academic metrics and applicants need to have better links with industry rather than an impressive publication record.

The assignment of FoR codes to projects reveals interesting patterns that may reflect both genuine interdisciplinarity and strategic behaviour. The results indicate that Linkage projects, which involve industry collaboration, exhibit significantly higher IDD values than Discovery projects. This means that industry engagement might foster multi- or inter-disciplinarity. Assigning FoR codes to projects by researchers at the time of grant application submission is not necessarily purely driven by the fields or topics covered in an application. Politics and tactics are involved as FoR codes assigned play a key role in the success of an application because it determines which ARC Panel of Experts will make the decision about a grant's success. There can be a bit of gaming involved in this. Past research on UK Research Assessment Exercise (RAE) showed that researchers engaged in a bit game-playing to influence the outcome of the assessment (Kelly and Burrows, 2012). In Australia, also a study on religious studies showed that researchers submit applications that are about religious studies but do not use the FoR code for that field (2204), instead use different codes that they perceive will increase their chance perhaps (Possamai et al., 2021). This might be more important in Discovery projects than in Linkage projects and it might be one of the reasons why Discovery projects are less likely to have fewer FoR codes.

The analysis reveals distinct patterns in how different disciplines engage with industry and how collaboration occurs across state boundaries. The dominance of engineering and physical sciences in Linkage projects, particularly in their connections with manufacturing and professional services sectors, suggests these fields have developed stronger industry engagement mechanisms or they are perhaps more inherently suitable for industry collaboration. Past research has alluded to the positive impacts of these grants, as Linkage projects have improved industry partners engagement with academics (Cassity and Ang, 2006). However, their long-term impact has not been adequately investigated.

The study has a few limitations. It relies on ARC data that lacks some elements (e.g. percentage value of each FoR code is not publicly available). I only examined the grants from three years and trends and patterns (for instance industry collaboration) might be different in other years. I also aggregated FoR codes and industry classification to higher level for pragmatic reasons as it is very difficult to meaningfully analyse and present hundreds of different detailed codes and classes in a short paper. I also did not consider the size of grants (i.e., the amount of funding) in the analysis.

In terms of implications for policy, the findings suggest a few areas for consideration including geographic concentration in NSW and Victoria that might indicate a need for initiative to broaden. The uneven distribution of grants across disciplines also needs examination. Here, I only analysed successful grants, and we do not know if under-funded areas do not apply or they simply have lower rate of success and there could be unintended biases in the evaluation process. The strong presence of public sector and professional service partners indicates potential opportunities to diversify industry engagement into other sectors.

Although the context of this study was Australian, its findings have broader implications for research funding policies globally. Many countries employ

competitive funding schemes that emphasise industry engagement, such as the Horizon Europe program or NSF Industry-University Cooperative Research Centers. The observed patterns in Australia - particularly the higher interdisciplinarity in industry-linked projects - suggest that similar schemes in other nations may also foster broader disciplinary integration. It would be useful to find out if this is the case in the USA and Europe, as it has implications for collaborative and interdisciplinary research.

Conclusion

This analysis of ARC Discovery and Linkage projects from 2023 to 2025 sheds some light on how research funding is distributed across disciplines, the extent of collaboration within and between states, and the level of industry engagement. The analysis reveals distinct patterns of interstate collaboration, with stronger connections among east coast institutions in both Linkage and Discovery projects, however, Linkage projects exhibit a more integrated collaboration pattern overall. A disproportionate number of Discovery grants are awarded to fields such as Engineering and Biological Sciences, while some fields (e.g., Creative Arts and Writing, Philosophy and Religious Studies) receive minimal funding. Linkage projects show a similar pattern but with stronger interdisciplinarity. The findings show that projects with industry linkage are more interdisciplinary, and more disciplines are involved in projects on average. The dominance of certain industry sectors, particularly in public administration and professional services, indicates potential room for diversifying industry engagement. Future research might investigate a few under-explored aspects of these projects including their long-term impact, the benefit of Linkage projects for industry partners, and factors that facilitate or hinder the participation of certain industry sectors to engage in Linkage projects.

Acknowledgement

I acknowledge the assistance of OpenAI's ChatGPT-4o in writing Python scripts for network analysis. I reviewed the scripts before use and checked their outputs for accuracy.

References

Alpaydın, U. A. R., & Fitjar, R. D. (2021). Proximity across the distant worlds of university–industry collaborations. *Papers in Regional Science, 100*(3), 689-712.

Australian Bureau of Statistics. (2006-revision-2.0). Australian and New Zealand Standard Industrial Classification (ANZSIC). ABS. https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-industrial-classification-anzsic/latest-release

Australian Bureau of Statistics. (2020). Australian and New Zealand Standard Research Classification (ANZSRC). ABS. https://www.abs.gov.au/statistics/classifications/australian-and-new-zealand-standard-research-classification-anzsrc/latest-release

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature, 534*(7609), 684-687.

Brooks, R. D., & Byrne, J. (2006). A citation analysis of ARC Discovery and Linkage grant investigators in economics and finance. *Applied Economics Letters, 13*(3), 141-146.

Cassity, E., & Ang, I. (2006). Humanities–industry partnerships and the 'Knowledge Society': the Australian experience. *Minerva, 44*(1), 47-63.

Cetindamar, D., Renando, C., Bliemel, M., & Klerk, S. D. (2024). The evolution of the Australian start-up and innovation ecosystem: Mapping policy developments, key actors, activities, and artefacts. *Science, Technology and Society, 29*(1), 13-33. doi:10.1177/09717218231201878

Cherney, A. (2015). Academic–industry collaborations and knowledge co-production in the social sciences. *Journal of Sociology, 51*(4), 1003-1016.

Clarke, K., Flanagan, J., & O'Neill, S. (2011). Winning ARC grants: comparing accounting with other commerce-related disciplines. *Accounting Research Journal, 24(*3), 213-244.

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics, 3*(3), 222-232.

Gläser, J., & Laudel, G. (2016). Governing science: How science policy shapes research content. European *Journal of Sociology/Archives Européennes de sociologie, 57*(1), 117-168.

Jackson, P., Mavi, R. K., Suseno, Y., & Standing, C. (2018). University-industry collaboration within the triple helix of innovation: The importance of mutuality. *Science and Public Policy, 45*(4), 553-564. doi:10.1093/SCIPOL/SCX083

Jackson, P., Runde, J., Dobson, P., & Richter, N. (2016). Identifying mechanisms influencing the emergence and success of innovation within national economies: a realist approach. *Policy Sciences, 49*, 233-256. doi:10.1007/s11077-015-9237-6

Jamali, H. R., Abbasi, A., & Bornmann, L. (2020). Research diversification and its relationship with publication counts and impact: a case study based on Australian professors. *Journal of Information Science, 46*(1), 131-144.

Lin, Y., Frey, C. B., & Wu, L. (2023). Remote collaboration fuses fewer breakthrough ideas. *Nature, 623*(7989), 987-991.

Maldonado, D., & Brooks, R. (2004). ARC Linkage projects and research-intensive organizations: are research-intensive organizations likely to participate? *Economic Papers, 23(*2), 175-188.

Ponds, R., Van Oort, F., & Frenken, K. (2007). The geographical and institutional proximity of research collaboration. *Papers in Regional Science, 86*(3), 423-444.

Possamai, A., Long, G., & Counted, V. (2021). An analysis of Australian Research Council's grants in religion. *Journal for the Academic Study of Religion*, (1).

Spre. (2023). *Education as an export for Australia*. http://www.spre.com.au/download/ExportsAustraliaStates2023.pdf

Tilbury, C., Bigby, C., & Hughes, M. (2020). Analysis of Australian Research Council grants awarded for social work projects 2008–2017. *Australian Social Work, 73*(1), 4-17.

Turner, G., & Brass, K. (2014). *Mapping the humanities, arts and social Sciences in Australia.* Canberra: Australian Academy of Humanities.

Universities Australia (2019, March 27). Australian Unis score top marks for world class research. https://universitiesaustralia.edu.au/media-item/australian-unis-score-top-marks-for-world-class-research/

Yegros-Yegros, A., Rafols, I., & D'este, P. (2015). Does interdisciplinary research lead to higher citation impact? The different effect of proximal and distal interdisciplinarity. *PloS ONE, 10*(8), e0135095.

https://doi.org/10.51408/issi2025_062

Investigating Information Propagation in Biomedical Literature through Citations: A Case Study

M. Janina Sarol[1], Halil Kilicoglu[2]

[1]*mjsarol@illinois.edu*
Informatics Programs, University of Illinois Urbana-Champaign, Champaign, Illinois (USA)

[2]*halil@illinois.edu*
School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, Illinois (USA)

Abstract

Scientific growth is iterative, with existing knowledge serving as the foundation for new discoveries. Citations serve as the primary channel for information propagation in science, shaping which ideas and findings persist in the literature and which do not. While natural language processing (NLP) is increasingly used in citation context analysis, it is underutilized in studies that examine the actual scientific content of citations. In this pilot study, we explored how NLP can be used to track the propagation of scientific findings by replicating a prior citation context study that relied on manual extraction. We compared two approaches: a traditional NLP pipeline (named entity recognition and relation extraction) and a generative large language model (LLM). We formulated a two-step automated pipeline: (1) extracting findings from a reference paper and (2) mapping citation contexts to the findings they reference. Our findings indicate that LLMs are superior to traditional NLP techniques in both steps of the pipeline. However, they are also more prone to errors, mapping citation contexts to findings they do not reference. While the two-step automated pipeline was effective, integrating manual annotation of findings with LLM-based mapping of citation contexts yields the best results. To our knowledge, this study is one of the first to explore how NLP, particularly LLMs, can be leveraged to track the flow of information in science. Future research should further evaluate the application of LLMs and other NLP techniques on a larger scale to assess their effectiveness in supporting citation-focused scientometric and informetric studies.

Introduction

Scientific progress is fundamentally cumulative, with each new discovery advancing upon prior knowledge. Citations serve as the primary means by which previous work is acknowledged, disseminated, and built upon (Cronin, 1984). They are the channels through which scientific information flows. As a result, analyzing citations can provide us with valuable insights into various aspects of science – the dynamics of scientific progress (Yang & Deng, 2024), influential research endeavors (Herrmannova et al., 2018), emerging trends (Schneider & Costas, 2017), and even gaps in current research (Farooq, 2017). It is therefore unsurprising that citations represent a core unit of analysis in science of science subfields such as bibliometrics, scientometrics, and informetrics.

Analyzing the scientific content within citation contexts could allow us to observe which ideas and findings continue to shape the literature, identify the most impactful discoveries within research domains, follow the emergence of new ideas, and track when and where scientific claims become generally accepted as facts. From an acknowledgment perspective, we can trace when and where scientific claims

originated and ensure proper recognition. Finally, it is important to note that not all citations are accurate (Jergas & Baethge, 2015), and unfounded information can make its way into the scientific record (Greenberg, 2009). With timely analysis of citation content, we could uncover and prevent misinformation from spreading in the scientific community.

A number of studies have explored tracing the information propagated from reference to citing articles. In early work, Cozzens (1985) investigated how citations to knowledge claims differed across two papers in Neuropharmacology and Sociology of Science. Most citations to the Neuropharmacology paper were about its methodology and findings, whereas a smaller percentage of citations to the Sociology of Science paper focused on its claims. Anderson and Lemken (2019) reviewed 1,400 citations to *Organizations*, a highly influential publication in management, and classified them into 7 thematic categories (classical organization theory, motivation and decisions, participation, conflicts, cognitive limits, routines and programs, and planning). Leng (2022) examined 343 papers citing a study about coronary heart disease and found that research communities tend to cite the findings most related to their communities.

While some citation context studies focus on the topical content of citations, they are often overshadowed by other higher-level analyses such as sentiment analysis and citation function classification. This is largely because these analyses are inherently easier to conduct, relying on classifying citations into a predefined set of categories that are applicable to all papers. For instance, for citation sentiment analysis, the goal is to classify citations into positive, negative, and neutral (Yu, 2013), while citation function classification categorizes citations based on their rhetorical purpose in the citing paper (Teufel et al., 2006). The relative ease of annotating data for these types of analyses has also contributed to the increasing availability of tools, particularly those leveraging natural language processing (NLP) techniques, which in turn makes it easier to conduct these types of studies at scale.

In contrast, (scientific) citation content varies from one paper to another, based on the source being cited, making it more challenging to develop generalizable NLP approaches tailored for this task. This is why content-focused citation studies typically involve manual analysis, which limits the number of citing publications researchers can feasibly examine. Coupled with the rapid growth of scientific literature and its citations, conducting generalizable content-focused citation studies is increasingly difficult. For example, as of February 2025, *Organizations* has accumulated over 40,000 citations according to Google Scholar. Extending the study by Anderson and Lemken (2019) to cover all citations would be a daunting task.

Although there is a lack of NLP-based approaches specifically developed for extracting and analyzing the scientific content of citations, various other NLP techniques may be useful for this task. Information extraction methods, in particular, can assist in automatically retrieving the scientific content of citations. By applying well-established tasks such as named entity recognition (NER) and relation extraction (RE), we can identify scientific concepts and their relationships mentioned in a reference paper and determine whether this information is cited by subsequent publications. For instance, Leng (2022) analyzed a paper by Paul et al. (1963), which

explored various factors associated with coronary heart disease. Leng (2022) identified 34 distinct findings, noting uneven citation distributions across these findings, with research communities typically citing the findings most relevant to their fields. It is possible to apply NER to determine the factors referenced in each citation, while RE could pinpoint which of the 34 findings were cited.

In this pilot study, we explore the potential of current NLP tools in tracking the flow of scientific information through citations. To showcase a real-world application, we aim to replicate the Leng (2022) citation context study. Our key research question is, "*How can we utilize NLP methods to effectively and efficiently track the propagation of information through citations?*" If full automation is not yet feasible, we assess which steps can be automated and which ones still require human intervention. We approach this by testing two methodologies: one that uses established NLP methods in NER and RE, and another that applies generative large language models (LLMs), which have recently attracted significant attention as a promising tool (Google DeepMind, 2024). Our study shows that NLP techniques, particularly LLMs, could help in understanding the flow of scientific information at scale while also suggesting that problems such as hallucinations need to be addressed to do this reliably.

Related Work

Tracking the propagation of information through citations is a well-explored research area, but it has primarily been approached from a network analysis perspective. For instance, della Briotta Parolo (2020) examined forward chains of citations to measure persistent influence, which describes how a paper impacts subsequent works in its citation chain, finding that publications linked to Nobel Prize winners have higher persistent influence. In contrast, Min et al. (2021) focused on the backward chain of citations, or references of references, to map the knowledge ancestry of papers. While these studies provide valuable insights, they overlook the actual content of citations, treating each citation as equally informative and important to the citing paper.

Automatically linking the citing text with the corresponding statements from reference articles has been explored, primarily for the task of scientific document summarization (Jaidka et al., 2016). Ou and Kim (2019) proposed similarity- and ranking-based methods for this task and suggested their use in conducting citation analysis studies. More recently, Sarol et al. (2024) connected citing texts with reference article statements to assess the accuracy of citations.

Methods

In this section, we give a thorough overview of the Leng (2022) citation context study, detail the specifics of our replication efforts, and describe the NLP solutions we evaluate.

The Leng Citation Context Study

Leng (2022) examined 343 publications that cited Paul et al. (1963), hereafter referred to as the *original study*, a prospective cohort study that examined several factors linked to coronary heart disease (CHD). Leng (2022) identified 34 different

findings from the original study. The citation contexts from each of these publications were manually extracted and classified based on the finding they referenced. 304 papers cited at least one finding, while 38 merely mentioned the original study without discussing any of its findings. One paper was found to cite incorrect information that did not appear in the original study. With its focus on citation context analysis to investigate how information from a single paper was propagated, Leng (2022) provides a strong foundation for our pilot study.
We categorized the findings discussed in Paul et al. (1963) into four sets of categories:

1) Association Relations

The original study found 15 factors associated with CHD: *cholesterol*, *blood pressure*, *coffee*, *smoking*, *body fatness*, *electrocardiogram findings* (particularly ST-segment or T-wave abnormalities), *somatotype* (primarily endomorphic dominance), *heart rate*, *chest discomfort*, *peptic ulcer*, *age*, *early death of father*, *chronic cough*, *shortness of breath*, and *arteriovenous nicking* (Paul et al., 1963). Although *diet* was not directly linked to *CHD*, a positive association was found between *diet* and *cholesterol levels*. There were 302 references to these association findings, with 195 papers citing at least one of them, representing over half (56.85%) of the citing papers. The association between *arteriovenous nicking* and *CHD*, however, was never cited.

2) Lack of Association

Paul et al. (1963) discovered 12 factors – *diet*, *alcohol*, *physical activity*, *body weight*, *job role*, *blood glucose*, *height*, *hemoglobin*, *gallbladder disease*, *lipoprotein lipase*, *non-paternal family history*, and *arcus senilis* – that appeared unrelated to CHD. These non-association findings were cited 124 times across 110 citing papers (32.07%). The non-associations between *CHD/family history* and *CHD/arcus senilis* received no citations.

3) Comparison

Paul et al. (1963) noted differences in dietary information based on the collection method. Dietary information collected using food diaries showed lower food intake than data from participant interviews. This finding was cited 7 times. 5 citing papers also compared the dietary intake between the original study participants and other population groups.

4) Other Findings

13 citing papers discussed the general incidence of *CHD* in the original study, without specifying its association to the factors. The seasonal fluctuations in serum cholesterol, seasonal fluctuations in blood pressure, and participation rate in the original study were cited by 6, 2, and 5 papers, respectively.
The citation counts of each finding are shown in Appendix Table 1. 231 papers cited a single finding, while 73 papers cited two or more findings. The most cited findings are the associations between *CHD* and *cholesterol* (85 citing papers), *blood pressure*

(57), and *coffee* (54). Additional analysis of the categorized citation contexts was conducted to determine which findings were cited together and how findings varied over time.
Finally, Leng (2022) constructed a citation network among the citing papers and partitioned the network into nine clusters, each representing a research community as inferred from the papers' titles. The network analysis revealed that (1) the distribution of findings highly varied, with no single finding being referenced by more than 25% of the papers, and (2) research communities primarily cited findings that aligned with their own research interests.

Replication

As the study focuses on the utility of NLP, our main goal is to automatically replicate the manual process of linking the content of each citation in a citing paper to the findings in the reference article. Specifically, given that the original study aimed to identify factors associated with *CHD*, we seek to identify citation contexts that referenced the association and lack of association findings. This replication will allow us to assess the feasibility of conducting such studies on a larger scale.
A total of 268 papers referenced at least one of these two groups of findings. We used the citation contexts extracted by Leng (2022), available in the supplementary material of this citation context study. Our automated process is as follows: we begin by extracting the findings from the original study, then classify the citing papers in accordance with those findings. This process simulates a scenario where a researcher fully relies on NLP, eliminating the need to manually read and extract findings from the reference paper.

Natural Language Processing Methods

We examined two methods: one that uses a combination of NER and RE, and another that solely relies on a large language model.

1) Named Entity Recognition and Relation Extraction

NER followed by RE is a common approach to extract knowledge from scientific literature in the form of concepts and their relationships, respectively. scispaCy is a Python library designed for processing biomedical and scientific texts (Neumann et al., 2019). It offers tools for biomedical NER, which is the NLP task of extracting biomedical concepts (entities) from unstructured text. Additionally, scispaCy supports entity linking, which normalizes different mentions of the same concept to standard identifiers in knowledge bases (French & McInnes, 2023). We mapped the concept mentions to their identifiers in the Unified Medical Language System (UMLS) (Bodenreider, 2004). For instance, the mentions *clinical coronary disease* and *coronary disease* both map to the same UMLS concept *coronary heart disease* (concept unique identifier: C0010068).
RE involves identifying related concepts based on the text and the nature of their relations. We performed relation extraction using the BERT-based model developed by Sarol et al. (2024) to identify associations and non-associations. This model was trained on the BioRED corpus (Luo et al., 2022) and extracts eight relation types:

association, positive correlation, negative correlation, binding, drug interaction, cotreatment, comparison, and conversion between six types of entities: diseases, chemicals, species, genes/proteins, mutations, and cell lines. Since the original study focused on associations, such as the link between *elevated cholesterol levels* and *CHD*, we broadened our definition of association to include both association and positive correlation predictions from the model. To determine lack of association, if scispaCy identified a pair of concepts (e.g., *CHD* and *diet*) in a citation context but the model did not detect a relation, we classified this as a lack of association.

2) Large Language Model

The NER + RE approach above is limited to some extent, as it can only consider the entity types included in UMLS, which, while extensive, is not exhaustive, and can only identify relations similar to those expressed in the BioRED corpus. Large language models have been shown to be capable of handling tasks they were not specifically trained on (Yang et al., 2024), making them a promising approach for this study. We designed two prompts: one to extract the findings from the original study and another to determine which findings were referenced in each citation context. In the first prompt, rather than instructing the LLM to identify concepts and their relations, we directly prompted the LLM to extract the original study's findings. The second prompt was applied individually to each citation context. We used Google Gemini 1.5 Pro as the LLM for this study, as it has demonstrated strong performance on long context documents (Google DeepMind, 2024), which makes it appropriate for processing scientific articles.

Table 1 shows the prompt used for the first step, with the input text truncated for readability. The input text contains the full text of the original study.

Table 1. Prompt for Identifying Findings in the Original Study.

Instruction	*The text below is a research publication. Please extract and summarize all the findings of this paper and present them in a structured JSON format. Ensure that each finding is concise, clearly worded, and reflects the main conclusions of the study.*
Input Text	*A Longitudinal Study of Coronary Heart Disease* *SINCE the Fall of 1957, a long-term study of coronary heart disease has been in progress at the Hawthorne Works of the Western Electric Company in Chicago under the auspices of the University of Illinois College of Medicine and Presbyterian-St. Luke's Hospital. The study was undertaken in the belief that coronary heart disease was a disease resulting from the interplay of multiple factors and that there was need to delineate these factors further…*

Figure 1 illustrates the JSON format of the output produced by the given prompt.

```
{
    "study_design": {
        "type": "Longitudinal",
        "duration": "4 years 5 months",
        "population": "1989 men aged 40-55",
        "location": "Hawthorne Works of the Western Electric Company, Chicago",
        "method": "Annual interviews and examinations"
    },
    "findings": [
        {
            "factor": "New Coronary Cases",
            "finding": "88 cases of coronary heart disease developed..."
        },
        {
            "factor": "Family History (Parental Longevity)",
            "finding": "No significant difference between coronary and non-coronary groups..."
        }
    ]
}
```


Figure 1. Output of the Prompt for Identifying Findings in the Original Study.

We constructed the second prompt based on the output of the first prompt. Table 2 shows an example prompt, which contains the instruction, the JSON-formatted list of findings obtained from the first prompt, and the citation context.

Table 2. Example Prompt for Identifying Cited Findings.

Instruction	*The JSON text below lists the findings of a reference paper. Each finding is described in the 'finding' field, with its shorthand provided in the 'factor' field. The text enclosed in $CITATION$ contains a citation to this reference paper.* *Identify which findings from the JSON are referenced in the citation text. A finding is considered cited if the information it conveys is consistent with the text in the 'finding' field.* *Output only the 'factor' values of the relevant findings as a comma-separated list. If no findings are cited, return an empty string.*
List of Findings	*[* *{* *"factor": "New Coronary Cases",* *"finding": "88 cases of coronary heart disease developed..."* *},* *{* *"factor": "Family History (Parental Longevity)",* *"finding": "No significant difference between coronary and non-coronary groups..."* *},* *...* *]*
Input Text	*$CITATION$* *A relationship between the serum cholesterol level and the relative risk of developing clinical coronary heart disease has been reported by many investigators [4, 6, 9, 11, 12, 14-16]". (p. 358)* *$CITATION$*

Evaluation

Recall served as our main evaluation metric for this study, as our goal was to determine if NLP could capture the same data as the manual approach. For each

finding, we calculated the proportion of citing publications correctly identified by the NLP methods. Precision was less suitable, particularly in the NER + RE approach, since it may extract valid biomedical concepts that were not part of the cited findings.

Results

Table 3 presents a comparison of the results of the two NLP methods. Overall, the LLM-based pipeline outperformed the traditional NLP approach. It identified 80% of the total citations to the findings, while the NER+RE approach only succeeded in mapping 23%. Further, out of the 268 citation contexts, the LLM correctly found all cited findings in 184 (69%) citation contexts compared to just 47 (18%) for the NER+RE method. The LLM successfully extracted 26 out of 28 findings from the original study (93%), whereas the NER+RE approach managed to retrieve only 16 (54%). We also examined the scenario in which findings are manually extracted (data was collected from the Leng (2022) study's supplementary material), automating only the process of mapping each citation context to the corresponding finding. The NER+RE approach had similar performance to the full 2-step pipeline, but the LLM method yielded better results when given manually annotated findings. A detailed list of recall results for each finding is available in Appendix Table 2.

Table 3. Summary of Results.

Task	*NLP Method*	*Recall*
Full Pipeline	NER + RE	23%
	LLM	80%
Step 1 Only: Extracting Findings from Original Study	NER + RE	57%
	LLM	93%
Step 2 Only: Mapping Findings to Citation Contexts	NER + RE	23%
	LLM	86%

Named Entity Recognition and Relation Extraction

Out of the 28 key concepts that scispaCy was tasked with identifying from the original study – *CHD* and the 27 studied factors – only one factor, *early death of father*, was not recognized. This factor does not correspond to a single UMLS concept. We found that despite the entity linking capabilities of scispaCy, manual entity linking was still necessary, as concepts in the original study could further be mapped to multiple UMLS concepts. For instance, the term *coronary heart disease* was used loosely in the original study, covering related concepts such as *angina pectoris* and *myocardial infarction*. Thus, we had to add these UMLS concepts to ensure that references to *CHD* were properly covered. The complete mapping is provided in Appendix Table 1. We used the UMLS identifiers to identify the concepts mentioned in the citation contexts.

The BERT-based model successfully extracted 4 of the 16 association relations from the original study, correctly linking *CHD* to *cholesterol*, *blood pressure*, *smoking*,

and *coffee*, coincidentally the four most cited findings. However, it erroneously detected an association between *CHD* and *hemoglobin*. The model also incorrectly mapped 8 findings to citation contexts that did not mention them – for example, the association between *CHD* and *height* was linked to a citation context that discussed the relationship between *CHD* and *cholesterol*.

Large Language Model

The LLM found 28 total findings (shown in Appendix Table 3). The lack of association between *CHD* and both *height* and *weight* were combined into a single finding. One of the findings is about the general incidence of *CHD* and another about the relation between *CHD* and *perceived tension*, which was not on the list of findings extracted by Leng (2022). However, it is indeed reported in the original study. All findings extracted by the LLM are accurate; the LLM did not hallucinate any findings. The LLM identified 14 association and all 12 lack of association findings. It failed to identify the associations between *diet* and *cholesterol*, and *CHD* and *age*. The LLM incorrectly attributed 46 incorrect findings to citation contexts that did not discuss them.

Discussion

In this pilot study, we attempted to replicate a study that focused on citation context analysis to understand how information is propagated from one article to others using automated methods. Our results show that LLMs are superior to more traditional information extraction methods in linking findings from a reference article to their citations.

The Need for Human Intervention

We note that in both methods, we still needed human intervention to complete the tasks. For the NER+RE pipeline, we needed to map *CHD* and each of the 27 factors to UMLS concepts, and this mapping was also limited to the UMLS concepts extracted from the original study. As a result, any concept that was not extracted from the original study, even if correctly identified in the citation contexts, was not included in the mapping. We found several UMLS concepts in the citation contexts that were consistent with those in the original study but were not extracted by scispaCy from the original study. For example, *cholesterol* and *alcohol*, both of which have 5% recall, had more than half of citation contexts containing mappings to *Serum cholesterol measurement* (C0587184) and *Alcohol consumption* (C0001948), respectively. Including these terms in the list of allowed UMLS concepts would raise their recall values to greater than 50%. Not only is there a need to manually map related UMLS concepts within the original study, but there is also a need to review the UMLS concepts extracted from the citation contexts, which may be an infeasible task to perform at scale. In contrast, the LLM only required minor human intervention, primarily for identifying the JSON format produced by the first prompt.

While the two-step automated pipeline using an LLM was shown to be effective, using the manually annotated findings with LLM-based mapping of citation contexts

yielded the best results. Of particular note was the boost in results related to *diet* when the manual annotations were used instead of the automatically extracted findings: recall of citation contexts citing findings on the association of *diet* and *cholesterol* and the lack of association of *CHD* and *diet* increased from 0% to 71% and 49% to 73%, respectively. This suggests that the most effective process still requires human intervention.

Human vs Machine Annotation

While the LLM yielded better recall performance, it also made more errors. We manually examined the 46 erroneous citation context-finding mappings and found that 8 were consistent with the citation context (i.e., they were manual annotation errors), 12 resulted from the extra text in the citation context (citation contexts were on a sentence level), and the remaining 26 were incorrect mappings by the LLM. Examples of each case are shown in Table 4. For the 8 incorrect manual annotations, 5 involved confusion between *job role* and *physical activity*. In the original study, *job role* referred to physical activity on the job, while *physical activity* referred to physical activity off the job.

Our analysis demonstrated that manual annotations are not consistently accurate, indicating the potential value of a hybrid approach that integrates both human and machine annotations. LLMs can either serve to supplement and double-check manual annotations or be regarded as independent annotators. However, it remains essential that humans perform the final verification and thoroughly review all annotation outputs.

Table 4. Examples of Erroneous Citation Context-Finding Mappings by the LLM.

Case	*Citation Context*	*CHD-Associated Factor*
Incorrect Manual Annotation	*However, studies are not entirely consistent, and a number of US long-term studies of initially healthy men have failed to show a relationship between incidence of ischemic heart disease and occupational activity [25-28]*	Manual: *physical activity* LLM: *job role* (no association)
Extra Text from the Citation Context	*Cigarette smoking is well established as a CHD-risk factor [17, 18], and caffeine intake has been incriminated recently [**19**]*	Manual: *coffee* LLM: *smoking*
Incorrect LLM Annotation	*Paul et al. [**30**] demonstrated a significant correlation between coffee consumption and the later development of coronary disease, although serum cholesterol levels were normal.*	Manual: *coffee* LLM: *cholesterol*

A Fully Automated Process

While our study aimed to replicate the manual mapping of citation contexts to referenced findings, arguably the most time-consuming part of the study, we skipped some necessary steps for full automation. A full end-to-end pipeline would include automated collection of the full texts of the original study and the citing papers, as well as the citation contexts pertaining to the original study. Both steps are non-

trivial. We initially tried collecting the list of citing papers automatically but failed to find most papers. We resorted to manually collecting the PDFs of citing papers and found that conversion from PDF to text is also an issue, especially since the citing papers are older documents published from 1963-1984. Future work should consider automating an end-to-end pipeline, which would be of most benefit to the scientometrics and informetrics communities.

Replicability to Other Publications

We note that Paul et al. (1963) is a short paper, and its findings are presented as section headers, making it a relatively easy case for a citation context study that tracks the dissemination of information. We considered it for this study, since the Leng (2022) study could be used as a proxy for ground truth. While the approach may yield weaker results on a more complex paper, this study demonstrates the potential for (semi-)automated approaches. Future work could consider the construction of a larger dataset that can be used for evaluation and possibly for training or fine-tuning NLP models, including LLMs.

Conclusion

We examined the potential of NLP in tracking the propagation of scientific findings through citations by replicating a citation context study that relied on manual extraction and assessing the advantages and shortcomings of two approaches: an NER and RE pipeline, and an LLM. LLMs outperformed the traditional NLP methods in both extracting findings from the original study and mapping citation contexts to their referenced findings. Our results suggest that LLMs might be an effective tool for analyzing the propagation of information in science. In the future, we plan to evaluate additional NLP tools and LLMs (including open-weight models) and refine this approach to apply it to other similar citation context studies to better assess its generalizability.

References

Anderson, M. H., & Lemken, R. K. (2019). An empirical assessment of the influence of March and Simon's Organizations: the realized contribution and unfulfilled promise of a masterpiece. *Journal of Management Studies, 56*(8), 1537–1569. https://doi.org/10.1111/joms.12527

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research, 32*(suppl_1), D267-D270. https://doi.org/10.1093/nar/gkh061

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

Cozzens, S. E. (1985). Comparing the sciences: citation context analysis of papers from neuropharmacology and the sociology of science. *Social Studies of Science, 15*(1), 127–153. https://doi.org/10.1177/030631285015001005

della Briotta Parolo, P., Kujala, R., Kaski, K., & Kivelä, M. (2020). Tracking the cumulative knowledge spreading in a comprehensive citation network. *Physical Review Research, 2*(1), 013181. https://doi.org/10.1103/PhysRevResearch.2.013181

Farooq, R. (2017). A framework for identifying research gap in social sciences: Evidence from the past. *IUP Journal of Management Research*, *16*(4), 66-75.

French, E., & McInnes, B. T. (2023). An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics, 137*, 104252. https://doi.org/10.1016/j.jbi.2022.104252

Google DeepMind. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens*. arXiv. https://arxiv.org/abs/2403.05530

Greenberg S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ, 339*, b2680. https://doi.org/10.1136/bmj.b2680

Herrmannova, D., Patton, R. M., Knoth, P., & Stahl, C. G. (2018). Do citations and readership identify seminal publications?. *Scientometrics*, *115*(1), 239-262. https://doi.org/10.1007/s11192-018-2669-y

Jaidka, K., Chandrasekaran, M. K., Rustagi, S., & Kan, M.-Y. (2016). Overview of the CL-SciSumm 2016 Shared Task. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)* (pp. 93–102).

Jergas, H., & Baethge, C. (2015). Quotation accuracy in medical journal articles—a systematic review and meta-analysis. *PeerJ, 3*, e1364. https://doi.org/10.7717/peerj.1364

Leng, R. I. (2022). Diversity in citations to a single study: A citation context network analysis of how evidence from a prospective cohort study was cited. *Quantitative Science Studies, 2*(4), 1216–1245. MIT Press. https://doi.org/10.1162/qss_a_00154

Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C. N., & Lu, Z. (2022). BioRED: A rich biomedical relation extraction dataset. *Briefings in Bioinformatics, 23*(5), bbac282. https://doi.org/10.1093/bib/bbac282

Min, C., Xu, J., Han, T., & Bu, Y. (2021). References of references: How far is the knowledge ancestry. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 262-265). IEEE. https://doi.org/10.1109/JCDL52503.2021.00079

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). SciSpaCy: Fast and robust models for biomedical natural language processing. *CoRR, abs/1902.07669*. https://doi.org/10.48550/arXiv.1902.07669

Ou, S., & Kim, H. (2019). Identification of citation and cited texts for fine-grained citation content analysis. *Proceedings of the Association for Information Science and Technology*, *56*(1), 740-741. https://doi.org/10.1002/pra2.156

Paul, O., Lepper, M. H., Phelan, W. H., Dupertuis, G. W., Macmillan, A., McKean, H., & Park, H. (1963). A longitudinal study of coronary heart disease. *Circulation, 28*(1), 20–31. https://doi.org/10.1161/01.cir.28.1.20

Sarol, M. J., Hong, G., Guerra, E., & Kilicoglu, H. (2024). Integrating deep learning architectures for enhanced biomedical relation extraction: a pipeline approach. *Database, 2024*, baae079. https://doi.org/10.1093/database/baae079

Schneider, J. W., & Costas, R. (2017). Identifying potential "breakthrough" publications using refined citation analyses: Three related explorative approaches. *Journal of the Association for Information Science and Technology*, *68*(3), 709-723. https://doi.org/10.1002/asi.23695

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural*

Language Processing (pp. 103–110). Association for Computational Linguistics. https://aclanthology.org/W06-1613

Yang, A. J., & Deng, S. (2024). Dynamic patterns of the disruptive and consolidating knowledge flows in Nobel-winning scientific breakthroughs. *Quantitative Science Studies*, *5*(4), 1070–1086. https://doi.org/10.1162/qss_a_00323

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*, *18*(6), 1-32. https://doi.org/10.1145/3649506

Yu, B. (2013). Automated citation sentiment analysis: What can we learn from biomedical researchers. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1-9. https://doi.org/10.1002/meet.14505001084

Appendix

Table 1. Named Entity Recognition Results and the Mapping of Concepts to UMLS Identifiers.

Concept	*UMLS Concepts*	*Citations*	*Recall*
Coronary Heart Disease	Coronary heart disease (C0010068) Coronary Arteriosclerosis (C0010054) Angina Pectoris (C0002962) Myocardial Infarction (C0027051)	264	62%
Cholesterol	Blood cholesterol (C0518017) Cholesterol measurement test (C0201950) Hypercholesterolemia (C0020443)	90	21%
Blood Pressure	Blood Pressure (C0005823) Systemic arterial pressure (C1272641) Diastolic blood pressure (C0428883) Hypertensive disease (C0020538)	57	86%
Coffee	Coffee (C0009237)	54	96%
Diet	Diet (C0012155) Eating (C0013470) fat intake (C0489488) salt intake (C0489767)	48	67%
Smoking	Smoking Habit (C4505437) Tobacco (C0040329) Cigar smoker (C0337666) Pipe Smoking (C4316784) Cigarette (C0677453) Cigarette smoke (substance) (C0239059)	43	72%
Body Fatness	Skinfold Thickness (C0037302) Skin-fold thickness (finding) (C0424680) Triceps skin fold thickness (observable entity) (C0518022)	29	24%
Physical Activity	Physically active (C0556453) Exercise (C0015259)	28	68%
Alcohol	Alcoholic Beverages (C0001967)	22	5%
Body Weight	Body Weight (C0005910) Weight Gain (C0043094)	13	54%
Electrocardiogram	Electrocardiogram (C0013798) Electrocardiogram finding (C0438154) Electrocardiographic changes (C0855329) Anatomical segmentation (C0441635) Abnormal T-wave (C1839341)	11	36%
Job Role	Occupations (C0028811)	7	71%
Blood Glucose	Blood Glucose (C0005802) Blood glucose measurement (C0392201)	6	67%
Somatotype	Somatotype (C0037669)	4	75%
Height	Height (C0489786)	3	100%
Heart Rate	Pulse Rate (C0232117)	3	33%
Peptic Ulcer	Peptic Ulcer (C0030920)	2	100%

Hemoglobin	Hemoglobin A measurement (C1281911) Chrysarobin (C0008721)*	2	100%
Age	Age (C0001779)	2	100%
Chest Discomfort	Chest discomfort (C0235710) Non-cardiac chest pain (C0476281)	2	50%
Chronic Cough	Chronic cough (C0010201)	1	100%
Gallbladder Disease	Gall Bladder Diseases (C0016977)	1	100%
Lipoprotein Lipase	LIPOPROTEIN LIPASE (C0023816)	1	100%
Shortness of Breath	Dyspnea (C0013404) Resting Dyspnea (C0743330)	1	100%
Early Death of Father		1	0%
Arteriovenous nicking	Retinal arteriovenous nicking (C1142247)	0	NA
Arcus Senilis	Arcus Senilis (C0003742)	0	NA
Family History	Family history (finding) (C0241889)	0	NA
TOTAL		695	59%

Table 2. Full. Pipeline Results: Mapping of Citing Papers to Findings *(L refers to lack of association).*

Relation	*Citations*	*Recall*			
		NER+RE		LLM	
		Step 2 Only	Full Pipeline	Step 2 Only	Full Pipeline
CHD/cholesterol	85	5%	5%	93%	96%
CHD/blood pressure	57	19%	19%	95%	96%
CHD/coffee	54	61%	61%	100%	93%
CHD/smoking	43	23%	23%	93%	98%
CHD/diet (L)	41	37%	37%	73%	49%
CHD/body fatness	29	0%	0%	59%	79%
CHD/physical activity (L)	28	32%	32%	71%	36%
CHD/alcohol (L)	22	5%	5%	91%	82%
CHD/body weight (L)	13	38%	38%	69%	38%
CHD/electrocardiogram	11	0%	0%	73%	73%
diet/cholesterol	7	0%	0%	71%	0%
CHD/job role (L)	7	0%	0%	86%	86%
CHD/blood glucose (L)	6	67%	67%	50%	33%
CHD/somatotype	4	0%	0%	100%	75%
CHD/height (L)	3	67%	67%	100%	100%
CHD/heart rate	3	0%	0%	67%	100%

CHD/age	2	50%	0%	100%	0%
CHD/chest discomfort	2	0%	0%	100%	100%
CHD/peptic ulcer	2	0%	0%	100%	100%
CHD/hemoglobin (L)	2	50%	50%	100%	100%
CHD/early death of father	1	0%	0%	100%	100%
CHD/chronic cough	1	0%	0%	100%	100%
CHD/shortness of breath	1	0%	0%	100%	100%
CHD/gallbladder disease (L)	1	100%	100%	100%	100%
CHD/lipoprotein lipase (L)	1	100%	100%	0%	0%
TOTAL	426	23%	23%	86%	80%

Table 3. Original Study Findings Identified using an LLM (Google Gemini 1.5 Pro).

Factor	*Finding*
New Coronary Cases	88 cases of coronary heart disease developed (47 angina pectoris, 28 myocardial infarction, 13 deaths). Approximately one new case per 100 men per year.
Family History (Parental Longevity)	No significant difference between coronary and non-coronary groups regarding parental age at death.
Family History (Paternal Age)	Fathers of non-coronary group lived 3.4 years longer on average than fathers of coronary group.
Prior Chest Discomfort	Significantly higher development of coronary disease in men reporting prior chest discomfort ($p < 0.001$).
Chronic Cough	Significantly more frequent in coronary group (25% vs 12%, $p < 0.001$).
Shortness of Breath	Significantly more frequent in coronary group (18% vs 11%, $p < 0.025$).
Peptic Ulcer History	Significantly more frequent in coronary group (18% vs 10%, $p < 0.025$).
Gallbladder Disease History	No significant relationship with coronary disease.
Height and Weight	No significant difference between groups.
Body Fatness (Skinfold Thickness)	Significantly greater in coronary group ($p < 0.025$ for triceps, $p < 0.01$ for scapular).
Somatotype	Suggests endomorphic dominance may be a factor, but requires larger sample size to confirm.
Resting Pulse Rate	While mean pulse rate was not significantly different, distribution differed significantly ($p = 0.025$), with coronary group having more extreme cases.
Blood Pressure	Positive relationship between elevated systolic ($p < 0.001$) and diastolic ($p < 0.05$) blood pressure and coronary disease.
Arcus Senilis	No relationship with coronary disease.

Arteriovenous Nicking (Fundi)	Significantly associated with coronary disease ($p < 0.005$).
Perceived Tension	No significant difference between groups.
Hemoglobin Levels	No significant difference between groups.
Cholesterol Levels	Significantly higher in coronary group ($p < 0.01$), with stepwise increases across angina, infarction, and death subgroups.
Lipoprotein Lipase	No relationship with coronary disease.
Blood Glucose (2-hour post 100g glucose)	No significant difference in mean levels, but significant heterogeneity of variance within groups, particularly angina.
Electrocardiogram (ST-segment/T-wave abnormalities)	Significantly associated with coronary disease ($p < 0.005$).
Job Type	No association with coronary disease.
Off-Job Physical Activity	No striking differences, though coronary group tended to report less sports participation.
Cigarette Smoking	Significant association with coronary disease ($p < 0.005$), with a stepwise increase in risk across angina, infarction, and death subgroups.
Diet (excluding coffee)	No significant association with coronary disease within the observed range of fat intake.
Coffee Consumption	Significant association with coronary disease ($p < 0.025$).
Alcohol Consumption	No association with coronary disease.

https://doi.org/10.51408/issi2025_002

Is There Life on Mars? Studying the Context of Uncertainty in Astrobiology

Iana Atanassova[1], Panggih Kusuma Ningrum[2], Nicolas Gutehrlé[3], Francis Lareau[4], Christophe Malaterre[5]

[1] *iana.atanassova@univ-fcomte.fr*
Université Marie et Louis Pasteur, CRIT, F-25000 Besançon (France)
Institut Universitaire de France (IUF), Paris (France)

[2] *panggih_kusuma.ningrum@univ-fcomte.fr*
Université Marie et Louis Pasteur, CRIT, F-25000 Besançon (France)

[3] *nicolas.gutehrle@univ-fcomte.fr*
Université Marie et Louis Pasteur, CRIT, F-25000 Besançon (France)

[4] *lareau.francis@courrier.uqam.ca*
Université de Sherbrooke, Dept of Philosophy and Applied Ethics, Sherbrooke (QC) J1K 2R1 (Canada)
Université du Québec à Montréal, Dept of Philosophy & CIRST, Montréal (QC) H3C 3P8 (Canada)

[5] *malaterre.christophe@uqam.ca*
Université du Québec à Montréal, Dept of Philosophy & CIRST, Montréal (QC) H3C 3P8 (Canada)

Abstract

While science is often portrayed as producing reliable knowledge, scientists tend to express caution about their claims, acknowledging nuances and doubt, all the more so in novel domains of research paved with unknowns. Uncertainty is an intrinsic aspect of scientific inquiry, particularly in recent fields such as astrobiology, which tackles numerous hard questions about the origin, evolution, and distribution of life on Earth and elsewhere. Mapping uncertainty in science matters for achieving a more accurate understanding of scientific knowledge. It also helps identify research domains at the frontiers of knowledge where unknowns are the most salient. In this article, we investigate the presence, distribution and context of uncertainty in the field of astrobiology. We analyze a comprehensive corpus of 3,698 research articles published in three major journals in the domain from 1968 to 2020. We use a linguistically motivated approach to identify expression of uncertainty in article full text. The corpus was further segmented into research topics using Latent Dirichlet Allocation (LDA) to investigate variations in uncertainty across subfields and over time. Our findings show that, while uncertainty has remained relatively stable over the 50 years covered by the corpus, constituting 20–25% of sentences on average, it varies significantly across research fields, highlighting areas where unknowns, doubts and speculations are more prevalent. The analysis also highlights relationships between expression of uncertainty and rhetorical structure. Indeed, higher uncertainty levels were observed in the beginning (introductions) and towards the end (conclusions) of research articles, while middle sections contained less uncertainty. Abstracts also tended to express a slightly higher level of uncertainty compared to main texts, especially with greater variability, suggesting their role in summarizing research and highlighting unknowns. To investigate the context of uncertainty, a lexical analysis was conducted to identify nouns most frequently associated with uncertainty within each topic. Terms such as “life,” “planet,” and “Mars” were found to be strongly associated with uncertainty. Conversely, terms related to experimentation and measurement, such as “sample” and “spectrum,” were linked to an absence of uncertainty, pointing at a dichotomy between speculative and evidence-based lines of inquiry. The findings contribute to a better understanding of

the field of astrobiology and exemplify the relevance of the proposed method to identify uncertainty-related concepts in corpora of full text publications. They also offer a foundation for future comparative studies across disciplines.

Introduction

Uncertainty is a foundational element of scientific inquiry, influencing every stage of the research process from formulating hypotheses to interpreting results. The construction of new scientific knowledge, by its nature, involves various degrees of uncertainty, arising from research hypotheses, methodological limitations, measurement errors, and the interpretative nature of scientific reasoning. Therefore, studying uncertainty is important to gain understanding on the mechanisms behind the construction of new knowledge. It also matters for better depicting the status of scientific knowledge and its variations in evidential support in different fields of inquiry. Indeed, scientific fields vary not just in terms of objects of investigation but also in terms of methods, maturity of research programs and social organization, thereby likely displaying noticeable nuances in terms of uncertainty. In the present contribution, we propose to investigate how uncertainty is expressed in the recent discipline of astrobiology.

Astrobiology is a multidisciplinary field encompassing areas such as prebiotic chemistry, systems chemistry, synthetic biology, atmospheric sciences, planetary sciences, and astronomy that emerged in the 1990s following early works in space life sciences and origin of life studies (Dick & Strick, 2004). Its unifying feature is the pursuit of hard and yet unresolved questions that require cross-disciplinary insights: What is life? How did it originate on Earth? Does it exist elsewhere in the universe? How might life evolve on a cosmic scale? According to the NASA Astrobiology Roadmap, astrobiology includes the search for habitable exoplanets, Mars exploration, studies of life's origins and early evolution, and research on life's adaptability on Earth and in space (Des Marais et al., 2003). Similarly, the AstRoMap European Astrobiology Roadmap frames astrobiology as the study of life's origin, evolution, and distribution within cosmic evolution, addressing habitability in the Solar System and beyond (Horneck et al., 2016). The broad scope of astrobiology, as well as its recent emergence and the relatively speculative nature of its research objectives make it a perfect target for assessing the expressions of uncertainty in scientific research.

To this aim, we propose to deploy a linguistically motivated approach for identifying and categorizing uncertainty onto a full-text corpus consisting of all research articles published in the three major astrobiology journals (from earliest publication date in 1968 up until 2020). This approach relies on the identification of specific terminological patterns in texts, thereby going beyond more classical analyses of uncertainty focusing on hedgers and boosters (Ningrum & Atanassova, 2024). Moreover, by using a topic model already fitted to the corpus (Malaterre & Lareau, 2023), the method makes it possible to investigate uncertainty over time and across different subfields of astrobiology, thereby revealing nuances across disciplinary contexts which are further examined by identifying discriminating terms associated

with uncertainty. Uncertainty is also analyzed as a function of document properties and rhetorical structure (e.g., text progression, length, abstracts vs. main texts).
In what follows, we first describe the corpus and the methods, and lay out the set of analyses we conducted. We then present the results and discuss them, notably considering directions for future work.

Corpus and Methods

The corpus consists of all full-text research articles of the three major astrobiology journals that had been assembled in (Malaterre & Lareau, 2023): *Astrobiology*, the *International Journal of Astrobiology (IJA)*, and *Origins of Life and Evolution of Biospheres (OLEB*, this latter journal being successively known as *Space Life Sciences* (1968-1973), *Origins of Life* (1974-1984), *Origins of Life and Evolution of the Biosphere* (1984-2004), and *Origins of Life and Evolution of Biospheres* (2005-2023); since 2024, the journal has been renamed *Discover Life*). Editorials, conference summaries, errata, discussion notes, and short articles (<4,000 characters) were removed so as to only keep research articles and their abstracts. This led to a corpus consisting of a total of 3,698 full-text articles, including 3,542 with abstracts, from 1968 to 2020, with a total of 705,636 sentences (out of which 26,355 correspond to abstracts and 679,281 to the main text of the articles).
The corpus underwent standard preprocessing, including cleaning, tokenization, and vectorization. For the topic model, part-of-speech (POS) tagging and lemmatization using the TreeTagger package (Schmid, 1994) were conducted, and only nouns, verbs, modals, adjectives, adverbs, proper nouns, and foreign words were retained. Stop words, words shorter than three characters, and those appearing in fewer than 20 documents were removed. A topic model with K=25 topics was applied to the text data using the LDA algorithm (Blei et al., 2003), following a manual review of models with various K values (Malaterre & Lareau, 2023). Topics were interpreted and named based on an examination of top words and top texts. To facilitate analysis, the topics were organized into clusters using Louvain community detection on a graph of topic-to-topic correlations in documents. In short, the topics can be grouped into four clusters: (A) focuses on life and survival, including microbial communities in extreme environments, space biology, spacecraft contamination, and conceptual studies like Fermi's paradox. (B) centers on the origins of life, exploring prebiotic chemistry, amino- and nucleic acids, molecular evolution, meteorite analyses, and definitions of life, including artificial life and protocells. (C) addresses planetary and astro-related topics, such as exoplanet habitability, planetary atmospheres, chirality, and energy-matter delivery from space. (D) investigates biosignatures and geological traces, covering Mars exploration, hydrothermal vents, biopaleontology, microfossils, and the search for water and habitability on other worlds.
This study adopts a linguistically motivated system developed by Ningrum et al. (2023) to detect scientific uncertainty in scholarly full texts that is built using the spaCy framework. The system was applied to the cleaned full-text corpus (including abstracts). The system uses a weakly supervised approach with a fine-grained annotation scheme to identify uncertainty expressions at the sentence level. Its pipeline integrates pattern matching, complex sentence analysis, and authorial

reference checks, leveraging a span-based method to pinpoint uncertainty in academic writings. For a detailed presentation of this system, see Ningrum & Atanassova (2023). Building on prior findings (Desclés et al., 2011; Ningrum et al., 2025) that emphasize the importance of multi-word phrases in identifying hedging and uncertainty, the system goes beyond simple linguistic markers, and also relies on linguistic patterns and features, such as part-of-speech (POS) tags, morphology, and syntactic dependencies. Unlike earlier studies that assume all uncertainty expressions must contain at least one uncertainty span (Medlock & Briscoe, 2007; Szarvas, 2008; Farkas et al., 2010), this approach treats uncertainty spans as trigger candidates that require further verification. The verification covers three main types of contextual shifts that can alter the true interpretation of scientific uncertainty expression: rebuttal expressions due to confirmation, rebuttal expressions due to neutral informative statements, and negation. Figure 1 shows several examples of sentences and annotations. Table 1 presents a description of the dataset with the number of documents for each topic, the total number of sentences and the number of sentences identified as containing uncertainty. We processed abstracts and main texts of articles separately.

1 - *"Evaluation* ***seems to be an unresolved matter*** *in…."* [Uncertainty]

2 - *"The* ***potential roles of X*** *in Y* ***remain speculative****."* [Uncertainty]

3 – *"...no evidence to support* ***this hypothesis****..."* [Absence of uncertainty due to negation]

4 – *"In order to test* ***whether X has a contribution to Y****, statistical analysis was employed...."*

[Absence of uncertainty with neutral informative statement]

5 – *"The high correlations scores confirm* ***hypothesis H3*****"**

[Absence of uncertainty due to confirmation]

Figure 1. Examples of sentences and annotations of uncertainty. In bold: expressions of uncertainty that trigger the analysis of the context, and underlined: contextual elements that are analyzed to confirm or refute the presence of uncertainty.

Table 1. Dataset description for the abstracts and article main texts: number of documents, total number of sentences and sentences containing uncertainty for each topic. Articles were assigned their dominant topic as determined by LDA.

Dominant topic	Abstracts			Article main texts		
	Nb of documents	Total nb of sentences	Nb of sent. with uncertainty	Nb of documents	Total nb of sentences	Nb of sent. with uncertainty
A-Bacteria-microbes	178	1,560	312	179	32,980	5,146
A-Cell-plant-animal	110	851	127	118	20,128	3,454
A-Life-civilization	156	970	363	177	33,305	10,636
A-Radiation-spore	178	1,490	228	178	29,373	4,018
A-Sample-mission	83	659	99	87	25,718	4,754
A-Science-mission	72	523	61	86	17,508	3,059
B-Amino-acid	91	539	152	95	14,188	3,215
B-Chemistry	282	1,713	378	291	37,655	6,963
B-Life-system	241	1,531	462	256	42,069	11,315
B-Organic-molecule	232	1,658	461	238	39,586	8,858
B-Protein-gene-RNA	129	941	272	136	20,194	5,096
B-Sample-chemistry	204	1,505	303	211	32,200	5,612
B-Surface-mineral-vesicle	153	1,143	292	156	26,208	5,300
C-Atmosphere	123	1,090	349	128	32,524	8,681
C-Chirality	126	691	184	140	18,376	4,382
C-Impact-particle	107	804	279	107	21,510	5,885
C-Planet-star	156	1,260	407	160	36,779	9,533
C-Value-model	124	889	273	124	23,493	6,056
D-Life-environment	110	844	274	125	28,392	8,970
D-Mars	97	797	233	101	24,169	6,344
D-Reaction-vents	134	1,058	312	145	24,535	6,167
D-Rock-sample	104	904	268	106	23,093	5,600
D-Spectra	125	1,018	195	125	24,501	4,145
D-Structure-geology	149	1,292	288	151	32,494	6,710
D-Water	78	625	222	78	18,303	5,188
Total	3,542	26,355	6,794	3,698	679,281	155,087

Analyses

Once identified, sentences with uncertainty were summed up for each article and analyzed across the corpus, especially to assess the influence of time and research domains (topics). Three main uncertainty measures were calculated: uncertainty as a function of time period U_p; uncertainty as a function of topic and time period $U_{j,p}$; and uncertainty as a function of topics U_j:

$$U_p = \frac{\sum_{d \in p} u_d / s_d}{N_p} \qquad U_{j,p} = \frac{\sum_{d \in p} u_d \times t_{j,d}}{\sum_{d \in p} s_d \times t_{j,d}} \qquad U_j = \frac{\sum_p U_{j,p}}{T}$$

where u_d is the number of sentences expressing uncertainty in a document d, s_d is the number of sentences in document d, N_p is the number of documents per time period p, $t_{j,d}$ is the % value of topic j in document d, T is the number of time periods (18 in the present case). This was done for abstracts only, for main texts only, and

for complete articles (abstracts and main texts jointly) to compare uncertainty expressed in abstracts and main texts.

Further analyses were conducted to examine uncertainty as a function of text length and text progression (excluding abstracts in both cases), the latter being defined as:

$$\text{for a given } g \in \{0, 1, \ldots, 100\},\ U_g = \frac{\sum_d u_{d,g}}{\sum_d s_{d,g}}$$

where $u_{d,g}$ is the number of sentences expressing uncertainty such that their relative position $h = rank(s) \times 100/s_d$, where $rank(s)$ is the absolute position of sentence s in document d (from 1 to s_d), is such that g is the entire number that is closest to h.

To investigate the context of uncertainty in astrobiology, we analyzed occurrences of nouns and proper nouns in the body of the articles in each identified topic. First, we extracted the most frequent nouns from sentences annotated with uncertainty for each topic, thus identifying key terms that frequently occur around expressions of uncertainty. To do this, we performed tokenization, POS-tagging and lemmatization of the dataset using the Python Natural Language Toolkit (NLTK). Articles were assigned their dominant topic as determined by the LDA topic model. Second, we calculated Precision, Recall, and F-measure scores for these nouns to assess their effectiveness in characterizing uncertainty. The F-measure is a metric used to evaluate the performance of a classification model, particularly in information retrieval and machine learning (Van Rijsbergen, 1979; Christen et al., 2024). In the context of classification and feature selection, it has been shown that the F-measure can be used to rank features with respect to their degree of association with a class (e.g., Alwidian et al., 2016; Lamirel et al., 2016). With this in mind, for a given term *t* and a set *S* of all the sentences of a given set *D* of documents, we define a class-association score $A_{c,t,S}$ that expresses the degree of association of *t* with a given class *c* in *S* as the harmonic mean:

$$A_{c,t,S} = 2 \times \frac{AP_{c,t,S} \times AR_{c,t,S}}{AP_{c,t,S} + AR_{c,t,S}}$$

where

$$AP_{c,t,S} = \frac{|\{s:\, s \in S \cap c, t \in s\}|}{|\{s:\, s \in S, t \in s\}|} \text{ and } AR_{c,t,S} = \frac{|\{s:\, s \in S \cap c, t \in s\}|}{|\{s:\, s \in S \cap c\}|},$$

s is a sentence, and class *c* is a class that can be either "presence of uncertainty" or "absence of uncertainty". This approach enabled us to identify and rank the nouns which were most strongly associated with the presence (or absence) of uncertainty within each topic, in order to better understand the primary subjects or concepts related to uncertainty discourse across the different topics in the dataset. We specifically examined this context for the top 5 and bottom 5 topics with respect to U_j. We also calculated class-association scores for the nouns of the top 5% and the bottom 5% of the articles (i.e.,u_d/s_d), in order to identify frequent concepts associated with uncertainty.

Results

The main results are summarized hereafter, with contrasting variations in uncertainty depending on context, notably research topics, but also rhetorical dimensions such as text length and text progression.

Uncertainty as a function of time

Results indicate a relatively stable expression of uncertainty over the fifty year span of the corpus, in the range of about 20 to 25% of document sentences (abstracts and main texts together) (Fig. 2). Percentage of uncertainty sentences can be as low as about 5%, while maximum uncertainty may reach about 50%, with some outlier documents scoring even above 60%. In any case, most of the corpus documents express a relatively high level of uncertainty which remains relatively unchanged over time, despite the introduction of two new journals in 2000 and underlying changes in topics (Malaterre & Lareau, 2023).

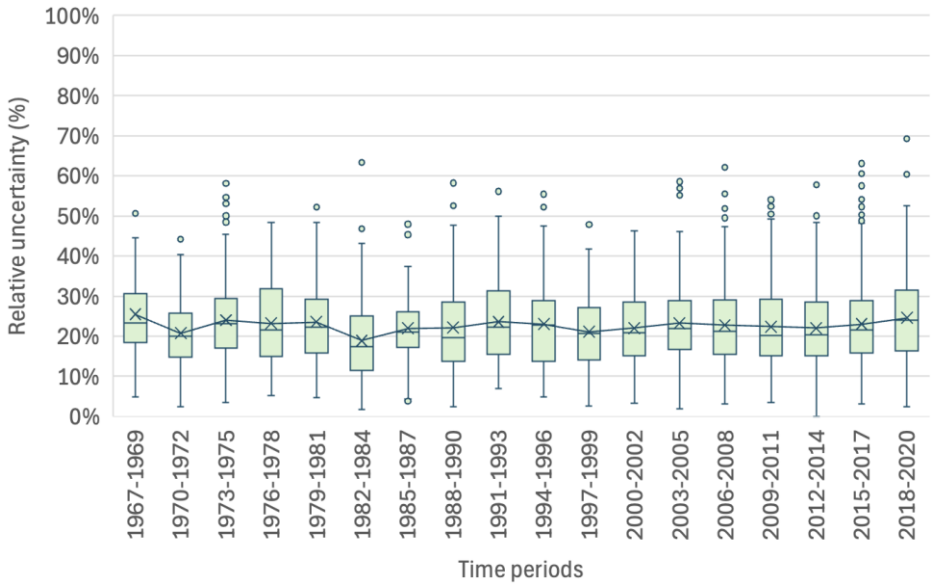


Figure 2. Evolution of share of uncertainty sentences in articles over time. For each time-period, boxplot of the distribution statistics of uncertainty percentage in research articles (abstract and main text); the line represents the evolution of average uncertainty U_p; dots are outlier articles.

Uncertainty per topic (excluding abstracts)

Analysis of uncertainty as a function of topic shows significant variation: while some topics express uncertainty in as few as about 15% of their attributed sentences, other topics have their share of uncertainty sentences well above 25% (Fig. 3). Among the five topics with least uncertainty, one finds three topics related to space microbiology ("A-Radiation-spore", "A-Bacteria-microbes", "A-Cell-plant-animal"), one to chemical analysis of rock samples ("B-Sample-chemistry), and one related to spectral analyses ("D-Spectra"). Among the five topics with the most uncertainty,

three concern life, its environment, whether alien civilization exists, what it means for a system to be alive ("D-Life-environment", "A-Life-Civilization", "B-Life-System") and two that concern astronomy, planetary systems in particular and impactors ("C-Planet-star", "C-Impact-Particle").

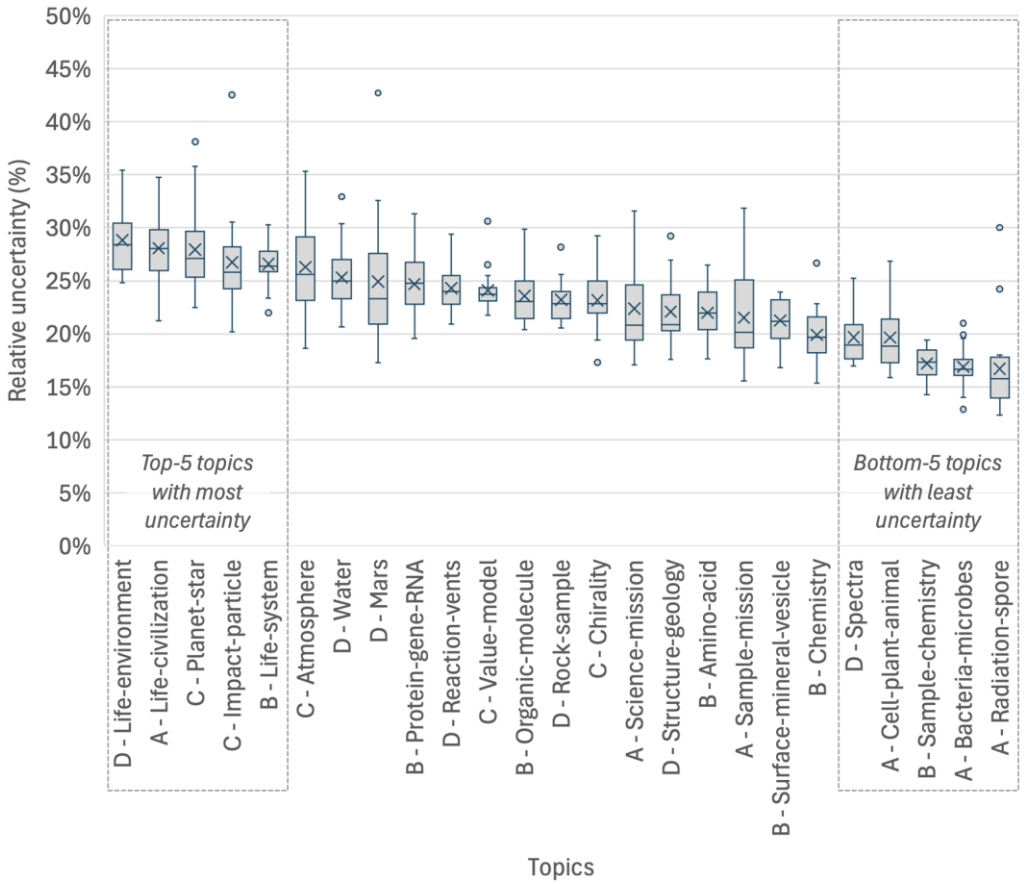


Figure 3. Share of uncertainty sentences as a function of topics. For each topic, boxplot of the distribution statistics of uncertainty % $U_{j,p}$ attributed to each of the 25 topics (for sentences in the main text only as abstract were not included in the topic model) across the 18 time-periods of the study; dots are outlier time-periods.

Context of uncertainty

To better understand the contexts of uncertainty, we examined the association scores with "uncertainty" and "absence of uncertainty" of all the nouns appearing in the corpus for various sets of documents (topic-related documents, outliers, and all corpus documents). Tables 2 and 3 present two different aspects of this analysis. As the different topics in the dataset contain various degrees of relative uncertainty (see Fig. 3), and the same phenomenon can be observed at the article level, we analyzed these association scores to identify the concepts that are most commonly related to the presence of uncertainty or to its absence. The highest association scores we observed on the dataset are about 0.36, thus the scores vary between 0 and 0.36. At

the article level, relative uncertainty in the main text varies between 1.43% and 69.81%. Due to this large interval, for the following analysis, we examined the outliers defined as the top 5% and the bottom 5% of articles, and compared them to the top-5 topics in terms of uncertainty, the bottom-5 topics, and to all corpus articles together.

Table 2 summarizes the highest association scores—above 0.1—with "presence of uncertainty" for the top-5 topics, top 5%articles, and all articles. We can observe, for example, that the noun "life" is highly related to the expression of uncertainty across all 5 topics, being at top position for three of them (D-Life-environment, A-Life-civilization, B-Life-system), and within the top-5 terms for the other two topics (C-Planet-star and C-Impact-particle). "Life" is also the highest ranking term among the top 5% articles expressing the most uncertainty, and across all articles of the corpus, but to a lesser extent. The nouns "planet" and "Earth" are present in all lists except for one topic (B-Life-system). Each topic presents its specificities, e.g. the uncertainty in D-Life-environment is prominently related to objects such as "environment", "Mars", "surface" and "condition" that do not appear in the other lists. Similarly, B-Life-system expresses uncertainty related to "molecule", "evolution" and "process" which are specific to that topic.

Table 3 lists the nouns that exhibit the highest association scores with "absence of uncertainty". We calculated these scores for the 5 topics that have the lowest relative uncertainty, for the bottom 5% articles in terms of relative uncertainty, and for all articles. Here, the term "sample" appears on the first or the second position for all lists except one topic (A-Cell-plant-animal). The lists that were obtained for the bottom 5% of articles and for all articles contain only one term ("sample") and no terms respectively. This can be explained by the much higher number of sentences without uncertainty compared to the number of sentences with uncertainty (about 4-fold, see Table 1); hence a much more diverse set of statements and vocabulary that cannot have high association scores with any specific noun.

Comparison between Tables 2 and 3 underscores insights on the types of research objects that are related to uncertainty within the different topics. Several terms in Table 3 appear related to experimentation and evidence-based research, e.g. "spectrum", "sample", "band", "cell", "experiment", "study", "acid", "temperature", "solution", "reaction", "spore". These nouns are strongly associated with the absence of uncertainty. In contrast, Table 2 indicates that uncertainty is expressed in relation with objects more prone to speculation or objects that are less directly observable or amenable to experimentation, e.g. "life", "planet", "Mars", "civilization", "star", "system", "atmosphere", "water", "evolution". Additionally, some objects can be related to the absence of uncertainty in some domains (e.g., "water" in the topic B-Sample-chemistry in Table 3), while being associated with the presence of uncertainty in other topics (D-Life-environment and C-Planet-star in Table 2). The term "time" is related to uncertainty for 3 topics in Table 2 but does not appear in Table 3. Similarly, "life" is prominently associated with uncertainty, while being absent from Table 3.

Table 2. Nouns with the highest association scores (above 0.1) with uncertainty for: the top 5 topics with highest relative uncertainty; the top 5% of articles with highest relative uncertainty; and all articles. Association scores are given in parentheses.

Topics with highest relative uncertainty										Top 5 % of articles		All articles	
D-Life-environment		A-Life-civilization		C-Planet-star		C-Impact-particle		B-Life-system					
life	(0.357)	life	(0.241)	planet	(0.352)	impact	(0.200)	life	(0.199)	life	(0.253)	life	(0.151)
Earth	(0.216)	planet	(0.122)	star	(0.212)	Earth	(0.178)	system	(0.188)	Earth	(0.164)	Earth	(0.116)
environment	(0.161)	Earth	(0.116)	Earth	(0.153)	life	(0.130)	molecule	(0.122)	planet	(0.136)	surface	(0.107)
planet	(0.160)	civilization	(0.110)	system	(0.152)	time	(0.127)	evolution	(0.119)	surface	(0.123)	planet	(0.101)
Mars	(0.139)	time	(0.106)	life	(0.132)	event	(0.122)	process	(0.109)	time	(0.116)		
surface	(0.129)			mass	(0.121)	planet	(0.109)			water	(0.112)		
water	(0.111)			atmosphere	(0.113)								
condition	(0.109)			time	(0.105)								
				water	(0.103)								

Table 3. Nouns with highest association scores (above 0.1) with *absence of uncertainty* for: the bottom 5 topics with lowest relative uncertainty; the bottom 5% of articles with lowest relative uncertainty; and all articles. Association scores are given in parentheses.

Topics with lowest relative uncertainty								Bottom 5 % of articles		All articles
D-Spectra		A-Cell-plant-animal		B-Sample-chemistry		A-Bacteria-microbes		A-Radiation-spore				
spectrum	(0.224)	cell	(0.183)	acid	(0.224)	sample	(0.194)	sample	(0.224)	sample	(0.141)	no terms
sample	(0.180)	experiment	(0.104)	sample	(0.200)	cell	(0.110)	radiation	(0.177)			
Raman	(0.153)	study	(0.103)	experiment	(0.140)			cell	(0.170)			
band	(0.136)			water	(0.111)			condition	(0.146)			
surface	(0.100)			solution	(0.107)			space	(0.142)			
				temperature	(0.106)			spore	(0.142)			
				reaction	(0.101)			experiment	(0.139)			
				compound	(0.101)			exposure	(0.134)			
								temperature	(0.107)			

Uncertainty in abstracts and in main texts

Uncertainty expressed in abstracts and in the body of articles tend to follow the same relatively stable pattern over time, though uncertainty in abstracts is usually a few points above uncertainty in the core of the texts. Note the higher variability of uncertainty expressed in abstracts, with most abstracts oscillating between 10% and 40% uncertainty, with minima at 0% and maxima or outliers oscillating between 80% and 100% uncertainty in some cases. The spread of uncertainty in the body of articles is much narrower, typically in between 15% and 25% of sentences expressing uncertainty.

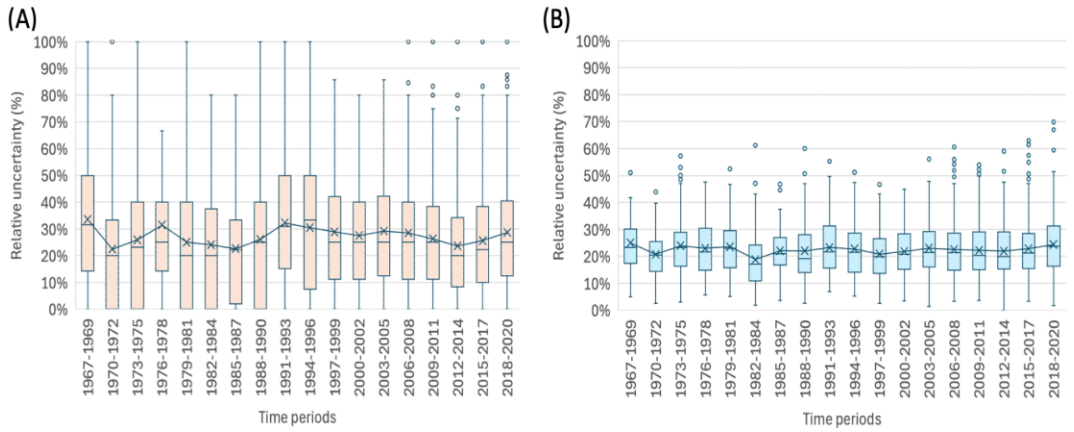


Figure 4. Comparison of uncertainty expressed in abstracts (A) and in the main portion of the corpus articles (B). Boxplot showing the distribution of document uncertainty ratio; line representing the evolution of average uncertainty per time-period.

Uncertainty as a function of text length

Analyzing text length as a function of uncertainty shows a lot of variability, though a noticeable trend seems to indicate that texts with either low or high uncertainty tend to be on the short side (around 100 sentences for texts with less than 10% uncertainty or more than 55%). On the other hand, texts with average uncertainty tend to be longer (about 200 sentences for texts with 20-30% uncertainty). This suggests that polarized texts in terms of uncertainty, exhibiting either a lot of doubt or a lot of conviction, tend to be on the shorter end.

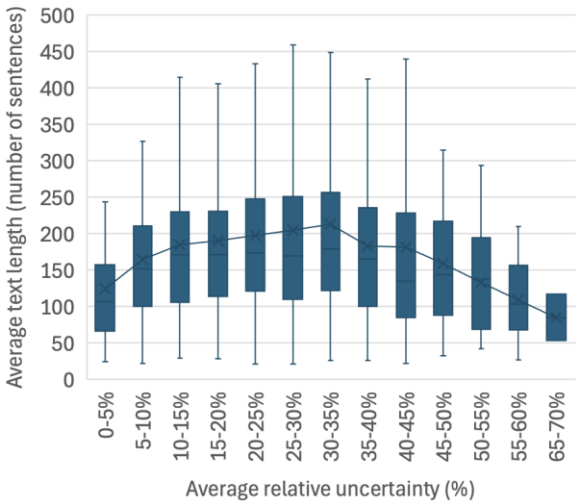


Figure 5. Document length as a function of uncertainty. For different intervals of uncertainty percentage in documents, boxplot of the distribution statistics of corresponding document length (total number of sentences in abstracts and main texts jointly).

Uncertainty as a function of text progression

Figure 6 shows the plot of the percentage of sentences that express uncertainty U_g with respect to their position in the text progression. Far from being constant throughout a text, uncertainty significantly fluctuates depending on text progression.

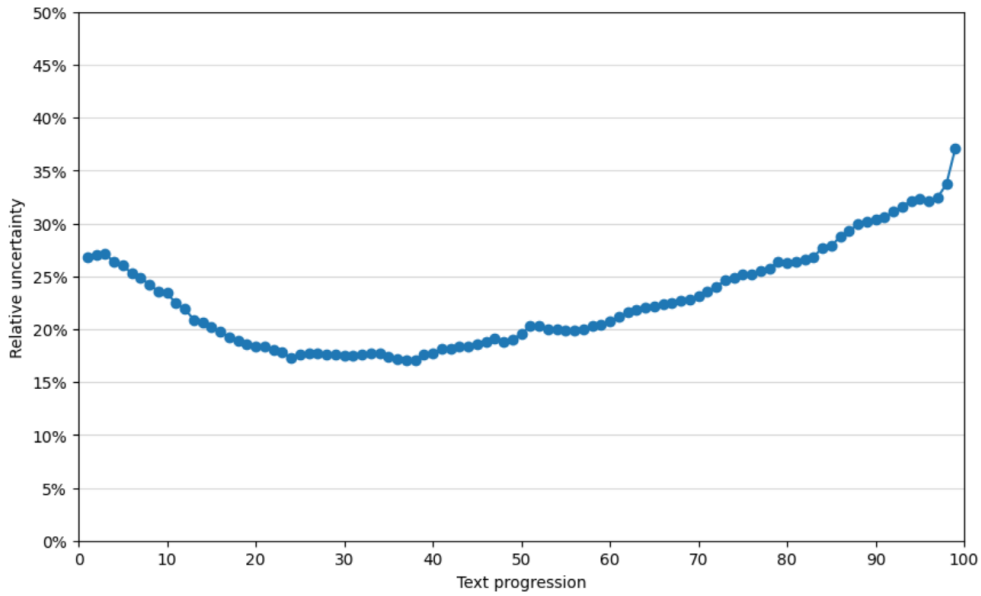


Figure 6. Relative distribution of uncertainty as a function of text progression (for main texts only, abstract excluded; 0 corresponds to text start; 100 to end).

The introductory portions of texts display a relatively high share of uncertainty, with as many as 27% of sentences expressing uncertainty. An even higher level of uncertainty is expressed in the concluding sections, with average uncertainty up to 37% at the end of texts. In between these two extremes, uncertainty levels are lowest between positions 20 and 40 of text progression. The IMRaD (Introduction, Methods, Results and Discussion) structure for articles is most usual in experimental sciences and commonly used in the journals in our dataset. Assuming such a structure for the majority of the corpus articles, uncertainty levels are rather high in the Introduction of the articles, at their lowest around the middle of the texts, i.e. in the Method and Result sections, and increase towards the final Discussion section.

Discussion

Our approach to annotating uncertainty, while effective, is not without limitations. The annotation relies on a set of nuanced linguistic rules to identify uncertainty, yielding an F-measure of 0.858 (Ningrum et al., 2025). While this performance is robust, it is not perfect and may introduce noise. Recent methodological improvements have been made, and further enhancements are planned.

The topic modeling approach which was used to identify research domains also has its constraints. We employed Latent Dirichlet Allocation (LDA) as it represents a well-established method, and fitted the model to K=25 topics so as to offer a

reasonable balance between granularity and research objectives, as addressed in prior work (Malaterre & Lareau, 2023). While providing nuanced topic probability distributions, this approach can impose certain limitations, for instance, the need for additionally crisp-assigning documents by assigning them to their dominant topics in some analyses.

Finally, the dataset itself presents limitations as it is confined to specific journals and time periods, reflecting a disciplinary focus on astrobiology. While this focus aligns with our objective of investigating uncertainty in that specific nascent multidisciplinary domain, extending the corpus to include articles from other journals using keyword-based retrieval could provide broader insights.

Our findings reveal that uncertainty in astrobiology research articles is relatively stable over time, both across the entire corpus and within specific topics. Contrary to initial expectations, uncertainty did not decrease over time, even as the field matured. While this challenges the hypothesis that uncertainty diminishes with disciplinary maturation, it remains possible that this trend could emerge at finer topic granularity than the 25 topics used in this study.

The corpus demonstrates relatively high levels of uncertainty, with on average about 20-25% of sentences in articles expressing uncertainty. This contrasts with previous studies that reported an average of 14% uncertainty in corpora from generalist and biomedical journals (Ningrum & Atanassova, 2024). Astrobiology thus occupies the higher end of the spectrum in terms of expressed uncertainty in the corpora examined so far. Moreover, individual articles vary widely, with some exhibiting as much as 60% uncertainty and others less than 10%. Investigating these extreme cases could yield valuable insights into the factors driving such variability.

One major finding is the significant variability in uncertainty across research topics. Certain topics express markedly more uncertainty, often linked to specific objects of inquiry. For example, particular nouns frequently associated with uncertainty suggest that the nature of the research object influences the level of expressed uncertainty. In the present study, the F-measure was used to identify most strongly associated nouns with specific groups of documents, yet the scores are low and furthermore the data is unbalanced; other measures, such as micro F-measure or TF-IDF at the cluster level (Grootendorst, 2022), could be used in future works. Future investigations should also explore in more detail whether epistemic properties—such as the difficulty of experimentation, observational challenges, or complexity—underlie this variability. One direction is to investigate the relationships between uncertainty and specific epistemic markers as defined in (Malaterre & Léonard, 2024). Additional sociological or cultural factors, such as differences in writing styles or practices, may also contribute and warrant further study.

Our analyses also highlight the interplay between uncertainty and the rhetorical structure of research articles. While there is no significant difference in average uncertainty between abstracts and main texts, abstracts exhibit greater variability in uncertainty levels. Interestingly, shorter texts tend to polarize in terms of uncertainty, displaying either very high or very low levels. Text progression emerges as a major variable influencing uncertainty. The introduction, discussion, and conclusion sections account for most instances of uncertainty, suggesting that these sections

function as rhetorical spaces for articulating doubt, speculation, and reflection. Comparative analyses across the IMRaD structure in different fields could further elucidate these patterns.

Conclusion

Deploying a linguistically motivated approach to identify complex terminological patterns expressing uncertainty in scientific articles, this study highlights the intricate dynamics of uncertainty within astrobiology research, offering insights into its relative stability over time, its variability across subdomains of research, and different facets of its rhetorical manifestations. Despite the field's maturation over the past fifty years, uncertainty remains prevalent, reflecting the challenges of investigating the origin on Earth and its possible presence elsewhere in the solar system and beyond. The variability of uncertainty across research domains—as captured with topic modeling —underscores different regimes of uncertainty possibly linked to specific objects of enquiry and their properties, and which will need to be further investigated. Lexical analysis identified nouns frequently linked to uncertainty, such as "life," "planet," and "Mars," contrasting with terms like "sample" and "spectrum," which reveal evidence-based inquiry. The analyses also highlight the relationship between uncertainty and the rhetorical structure of scientific articles. Higher uncertainty is found in introductions and conclusions, while middle sections contain less. Abstracts show slightly higher and more variable uncertainty, emphasizing their role in summarizing research and unknowns. These findings not only contribute to our understanding of the science of astrobiology and the uncertainty that pervades it, but also open pathways for comparative studies with other corpora and methodological refinements, notably to identify different types of uncertainties and further examine the epistemic context in which uncertainty is expressed. By extending these lines of enquiry, future research can further illuminate the nuanced role of uncertainty in scientific discourse.

Acknowledgments

C.M. acknowledges funding from Canada Social Sciences and Humanities Research Council (Grant 430-2018-00899) and Canada Research Chairs (CRC-950-230795). F.L. acknowledges funding from Canada Social Sciences and Humanities Research Council (Postdoctoral Fellowships 756-2024-0557). I.A., N.G. and P.K.N. acknowledge funding from the French ANR Project InSciM "Modelling Uncertainty in Science" (2021-2025) under grant number ANR-21-CE38-0003-01.

References

Alwidian, J., Hammo, B., & Obeid, N. (2016). Enhanced CBA algorithm based on apriori optimization and statistical ranking measure. *Proceeding of 28th International Business Information Management Association (IBIMA) Conference on Vision, 2020*, 4291–4306.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Christen, P., Hand, D. J., & Kirielle, N. (2024). A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Computing Surveys*, *56*(3), 1–24. https://doi.org/10.1145/3606367

Des Marais, D. J., Allamandola, L. J., Benner, S. A., Boss, A. P., Deamer, D., Falkowski, P. G., Farmer, J. D., Hedges, S. B., Jakosky, B. M., Knoll, A. H., Liskowsky, D. R., Meadows, V. S., Meyer, M. A., Pilcher, C. B., Nealson, K. H., Spormann, A. M., Trent, J. D., Turner, W. W., Woolf, N. J., & Yorke, H. W. (2003). The NASA Astrobiology Roadmap. *Astrobiology*, *3*(2), 219–235. https://doi.org/10.1089/153110703769016299

Desclés, J., Alrahabi, M., & Desclés, J.-P. (2011). BioExcom: Detection and Categorization of Speculative Sentences in Biomedical Literature. In Z. Vetulani (Ed.), *Human Language Technology. Challenges for Computer Science and Linguistics* (pp. 478–489). Springer. https://doi.org/10.1007/978-3-642-20095-3_44

Dick, S. J., & Strick, J. E. (2004). *The Living Universe NASA and the Development of Astrobiology*. Piscataway, NJ: Rutgers University Press.

Farkas, R., Vincze, V., Móra, G., Csirik, J., & Szarvas, G. (2010). The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning–Shared Task*, 1–12.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Horneck, G., Walter, N., Westall, F., Grenfell, J. L., Martin, W. F., Gomez, F., Leuko, S., Lee, N., Onofri, S., Tsiganis, K., Saladino, R., Pilat-Lohinger, E., Palomba, E., Harrison, J., Rull, F., Muller, C., Strazzulla, G., Brucato, J. R., Rettberg, P., & Capria, M. T. (2016). AstRoMap European Astrobiology Roadmap. *Astrobiology*, *16*(3), 201–243. https://doi.org/10.1089/ast.2015.1441

Lamirel, J.-C., Dugué, N., & Cuxac, P. (2016). New efficient clustering quality indexes. *2016 International Joint Conference on Neural Networks (IJCNN)*, 3649–3657.

Malaterre, C., & Lareau, F. (2023). The Emergence of Astrobiology: A Topic-Modeling Perspective. *Astrobiology*, *23*(5), 496–512. https://doi.org/10.1089/ast.2022.0122

Malaterre, C., & Léonard, M. (2024). Epistemic Markers in the Scientific Discourse. *Philosophy of Science*, *91*(1), 151–174. https://doi.org/10.1017/psa.2023.97

Medlock, B., & Briscoe, T. (2007). Weakly supervised learning for hedge classification in scientific literature. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 992–999.

Ningrum, P. K., & Atanassova, I. (2024). Annotation of scientific uncertainty using linguistic patterns. *Scientometrics*, *129*, 6261–6285. https://doi.org/10.1007/s11192-024-05009-z

Ningrum, P. K., Guterhlé, N., & Atanassova, I. (2025). Étudier l'incertitude dans les articles scientifiques: Mise en perspective d'une méthode linguistique. *Conférence Extraction et Gestion Des Connaissances (EGC)*.

Ningrum, P. K., Mayr, P., & Atanassova, I. (2023). UnScientify: Detecting Scientific Uncertainty in Scholarly Full Text. *Proceedings of ACM Conference (Workshop'23)*. EEKE2023, Santa Fe, New Mexico, USA.

Schmid, H. (1994). Part-of-speech tagging with neural networks. *Proceedings of the 15th Conference on Computational Linguistics-Volume 1*, 172–176. https://doi.org/10.3115/991886.991915

Szarvas, G. (2008). Hedge classification in biomedical texts with a weakly supervised selection of keywords. *Proceedings of Acl-08: HLT*, 281–289.

Van Rijsbergen, C. J. (1979). *Information Retrieval*. London: Butterworths and Co.

https://doi.org/10.51408/issi2025_054

Kazakhstani Scientific Collaboration with Post-Soviet Countries: Dynamics and Impact

Gulnaz Alibekova[1], Asselya Makanova[2]

[1]*galibekova77@gmail.com*, [2]*asselya.mak@gmail.com*
Institute of Economics of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Kazakhstan)

Abstract

This study explores the dynamics of international collaboration in scientific publishing by Kazakhstani authors, with a particular focus on co-authorships with scholars from post-Soviet countries. Drawing on bibliometric data from Scopus and SciVal, we analyze the evolution of Kazakhstan's scientific output in international collaborations over the 2011-2023 period. The results reveal notable trends in the geographical and institutional distribution of joint publications, shedding light on the relative impact of such collaborations on Kazakhstan's academic visibility.

A key finding of this study is the persistence and strengthening of scientific ties between Kazakhstan and other post-Soviet countries, despite the collapse of the Soviet Union over three decades ago. This trend appears to have been further reinforced by the introduction of indicator-based research evaluation systems and the adoption of policies aimed at the internationalization of science and higher education in post-Soviet countries, including Kazakhstan. As a result, Kazakhstani researchers increasingly engage in joint publications with colleagues from former Soviet republics. Notably, the volume of publications in collaboration with Russian authors remains the highest, although growing collaboration with researchers from Central Asia, particularly Uzbekistan, has also been observed.

In terms of publication quality, the study reveals that articles co-authored by Kazakhstani researchers with scholars from post-Soviet countries tend to be cited more frequently than those published with authors from other regions. Between 2014 and 2023, the average citation count of joint publications with post-Soviet colleagues was 19.87, compared to 13.52 for those co-authored with non-post-Soviet researchers. Moreover, publications with post-Soviet collaborators were more likely to appear in journals with lower impact-factor quartiles (Q3 and Q4), whereas those with non-post-Soviet co-authors were more frequently published in higher-impact journals (Q1 and Q2).

Introduction

Since joining the Bologna Declaration in 2010, Kazakhstan has pursued integration into the international scientific and educational community. In 2011, foundational regulatory documents were adopted, formalizing the transition from the Soviet model to Western standards for conducting research and implementing an indicator-based research evaluation system (Marina & Sterligov, 2021). Bibliometric analysis, which evaluates publication activity in international journals indexed in Scopus and Web of Science (WoS), became the primary method of assessment. Through this approach, the government initiated a policy aimed at internationalizing research in Kazakhstan (Moldashev et al., 2020).

One of the key indicators of integration into the global scientific community is the production of joint publications in international collaborations. Significant changes have been observed in this regard among Kazakhstani authors. For instance, the share of internationally co-authored publications in Scopus increased from 46.6% in 2011 to 53.2% in 2023.

As noted by Matveeva et al. (2023), in 1993, 19% of internationally co-authored publications by Kazakhstani researchers involved collaboration with post-Soviet scholars. According to Scopus data, in 1990–1991, Kazakhstani authors published 71 articles, 11 of which (15.5%) were written in international collaboration.

The shift in research priorities, particularly the adoption of internationalization policies in Kazakhstan and other post-Soviet countries, has led to the revival of previous scientific connections for co-authoring and publishing in international journals.

The purpose of this paper is to analyze the level of scientific collaboration between Kazakhstani researchers and their post-Soviet counterparts, 33 years after the dissolution of the Soviet Union, based on publications indexed in Scopus.

Methods

The data sources for analyzing the bibliometric indicators of publication activity by Kazakhstani authors in journals indexed in Scopus included the official website of www.scopus.com, as well as the analytical platform SciVal by Elsevier (www.scival.com). The primary data from these sources were collected on April 9, 2024, with additional data collected on July 28, 2024, and January 20, 2025. The choice of Scopus as the data source is due to its inclusion of a broader range of journals across all scientific disciplines (Mongeon & Paul-Hus, 2016), providing a more objective representation of international collaboration among authors from Kazakhstan.

To determine the total number of publications by Kazakhstani authors in Scopus, the advanced search query "AFFILCOUNTRY(Kazakhstan)" was used. The analysis considered all types of publications. To identify the number and share of publications by Kazakhstani authors in international collaboration in Scopus, the following approach was applied: using SciVal's "Explore" section, the publication date range was set to 2014–2023, and the "Country/Region" filter was applied to select "Kazakhstan." Subsequently, the "Collaborations" filter was used to isolate "International Collaborations," and post-Soviet countries (Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kyrgyzstan, Latvia, Lithuania, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine, and Uzbekistan) were selected under the "Country/Region" filter.

Data on institutional-level international collaboration were obtained via the SciVal search feature, where the names of Kazakhstani organizations in English were entered.

To calculate average citation rates, the number of publications with post-Soviet countries, the proportion of such publications, and analyze publication activity and citation counts by Kazakhstani institutions, the standard functionality of Excel was employed.

Research development in Kazakhstan: an overview

Kazakhstan has prioritized the internationalization of its scientific endeavors, aiming for deeper integration into the global academic community. This strategic shift aligns with broader efforts to modernize its research and education systems, reflecting

global standards and fostering collaboration with leading international institutions. The adoption of the *"Law on Science"* (MES, 2011a) in 2011 and the introduction of new regulatory acts in the field of science (MES, 2011b; MES, 2011c) solidified the transition from the Soviet model of scientific training to a Western framework. These changes were further driven by Kazakhstan's accession to the Bologna Declaration as part of the so-called "package of post-socialist reforms" (Kerimkulova & Kuzhabekova, 2017).

These legal reforms and the government's drive to establish a scientific system aligned with Western standards had a profound impact on the development of research activities in Kazakhstan. According to data from SciVal and the Scopus database (Elsevier), Kazakhstani authors published 50,973 articles in Scopus-indexed journals between 2011 and 2023. Given that the total number of publications by authors affiliated with Kazakhstan in the Scopus database across all years is 59,734, this means that 85.3% of all publications occurred during this period.

The most significant year-over-year growth in publication output followed the adoption of new regulatory acts in the field of science in 2011 (see Figure 1). While publication growth between 2011 and 2019 was characterized by fluctuations—marked by sharp increases and declines due to the turbulence of the transitional period—since 2020, the trend has become more stable.

The spikes in publication output during earlier years were partly driven by the effects of publication-focused policies, commonly referred to in the literature as the "publish or perish" phenomenon (Kurambayev & Freedman, 2021). This policy also had negative consequences, such as a rise in publications in predatory journals (Kudaibergenova et al., 2022; Marina & Sterligov, 2021). The growth in such publications was fueled by some researchers' attempts to meet formal requirements for publishing in international databases, often as a prerequisite for earning academic degrees, titles, grants, or points in institutional internal performance rankings (Yessirkepov et al., 2015). For some researchers, these requirements became a "game," prompting them to develop strategies to improve their chances of being published in appropriate journals (Moldashev et al., 2019) and achieve their professional goals.

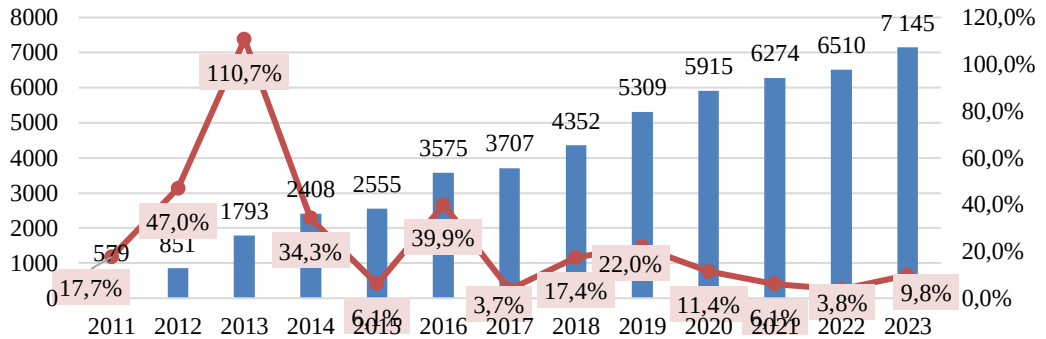


Figure 1. The number and annual growth rate (%, right axis) of publications by Kazakhstani authors in Scopus-indexed journals, 2011–2023.

For example, the relaxation of publication requirements for scholars in the social sciences and humanities in 2015 led some researchers to adopt a "gaming" strategy (Smagulov et al., 2018; Moldashev et al., 2020).

Thus, the "publish or perish" policy encourages researchers to publish in international journals but can also foster unethical practices (Kurambayev & Freedman, 2021). Such practices include publishing in predatory journals, relying on paper mills that provide publication services for international databases, or purchasing authorship slots in pre-written articles markets. For instance, through a single paper mill, Kazakhstani authors published 542 articles between 2019 and 2022 (Abalkina, 2023).

By 2020, the regulatory framework for scientific publications in international journals was harmonized, reflecting a unified scientific policy (Turginbayeva & Makanova, 2024). Detailed criteria were introduced for journals (percentile rankings in CiteScore, citation indices in WoS, and thematic alignment), authors (first or corresponding author), and articles (relevance to the journal's focus). These measures aimed to combat publications in questionable journals and discourage formalistic approaches to publishing.

The policy of internationalizing science has also encouraged active collaboration between Kazakhstani researchers and their international colleagues in co-authoring and publishing articles in reputable journals. However, a negative consequence of this process has been the "internationalization of unethical practices." For instance, Kazakhstani researchers frequently engage with foreign paper mills that offer services such as (a) publishing an author’s completed article, (b) selling authorship slots in pre-written articles, or (c) article writing and publishing in international journals. These schemes often involve collaboration with authors from other countries, primarily from the post-Soviet region. For example, the main clients of a Russian paper mill during 2019–2022 included researchers from Russia (2,715 articles), Kazakhstan (542 articles), and Ukraine (111 articles) (Abalkina, 2023).

Results

Publishing joint articles with foreign researchers is a key indicator of the integration of Kazakhstan's scientific community into the global research landscape. Furthermore, articles co-authored with international scholars positively impact citation metrics (Chankseliani et al., 2021).

The level of international collaboration (the proportion of articles co-authored with foreign researchers in Scopus-indexed journals) increased from 46.6% in 2011 to 53.2% in 2023. The lowest value was observed in 2013, at 35.4%, while the highest value was recorded in 2021, at 57.4% (Figure 2).

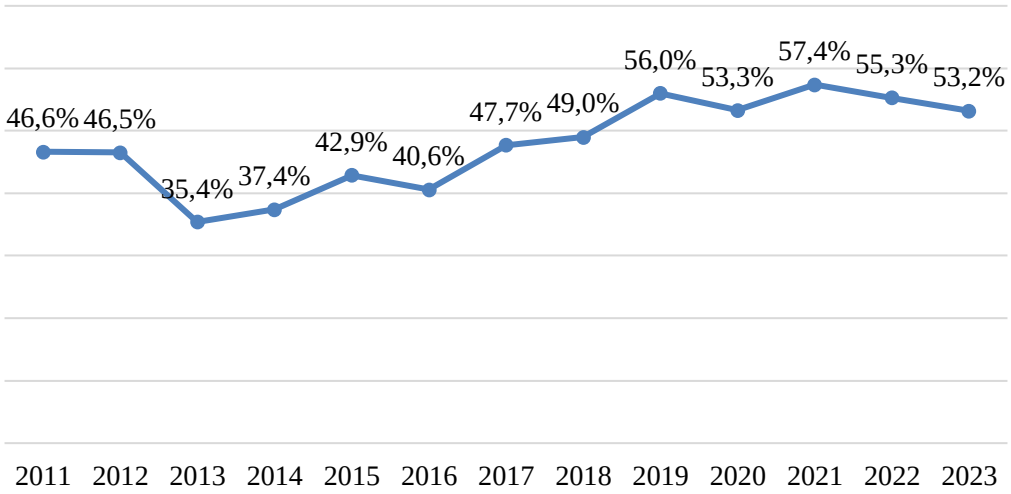


Figure 2. Proportion of publications by Kazakhstani authors in international collaboration in Scopus-indexed journals.

According to Scopus data, Kazakhstani authors collaborated with researchers from 68 countries in 2011, while this number increased to 174 in 2023. This indicates that the policy aimed at internationalizing Kazakhstani scientific research and integrating it into the global scientific community has been effective.

Since 2015, the top 10 countries with which Kazakhstani authors collaborate most frequently have remained relatively stable, with 9 leading countries consistently appearing across the years (Table 1). Throughout the analyzed period, the majority of articles authored by Kazakhstani researchers in Scopus-indexed journals have been co-authored with Russian scientists. This trend is attributed to the strong historical ties between Kazakhstan and Russia in the scientific domain, rooted in the Soviet era. These connections exist both on a personal level (e.g., university education, joint work experiences) and at the institutional level (e.g., collaborative educational programs and projects). This suggests that post-Soviet countries, including Kazakhstan, find it difficult to move away from empirical approaches that are still oriented toward Russia (Chankseliani, 2017).

Table 1. Top 10 countries collaborating with Kazakhstani authors in Scopus-indexed journals by year.

	2011	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*	*2022*	*2023*
Russia	113	132	210	301	351	557	719	774	1177	1293	1419	1291	1299
United States	64	100	102	151	146	224	260	349	322	336	363	433	462
Ukraine	15		48	47	92	124	178	181	321	243	323	321	315
China		29	33	38	53	102	108	142	255	235	363	357	353

United Kingdom	21	53	66	89	76	146	153	183	190	179	270	301	269
Poland		30	37	75	97	94	185	187	278	216	276	257	220
Germany	31	58	86	101	89	104	116	114	168	169	210	236	207
Turkey	11				75	72	97	115	141	133	237	253	284
Italy	14	33	32	50	54	77	100	125	139	140		208	184
India								95		136	206	205	229
France	16			39		62	61		122				
Saidi Arabia											200		
Japan	18	39	39	44	40								
Spain		36	29										
Pakistan		30											
Uzbekistan	10												

It is important to highlight that scientific collaborations between Kazakhstani authors and researchers from former Soviet Union countries remain dominant. Between 2011 and 2023, 51.03% of joint publications by authors affiliated with Kazakhstan were co-authored with researchers from these countries. This dynamic has remained relatively stable over the years: 53.3% of all Kazakhstani publications in international collaboration in 2011 and 50.6% in 2023 (Figure 3).

Remarkably, this stability has persisted despite a significant increase in the absolute number of publications in international collaboration, which grew from 259 in 2011 to 3,808 in 2023—an increase of 14.7 times.

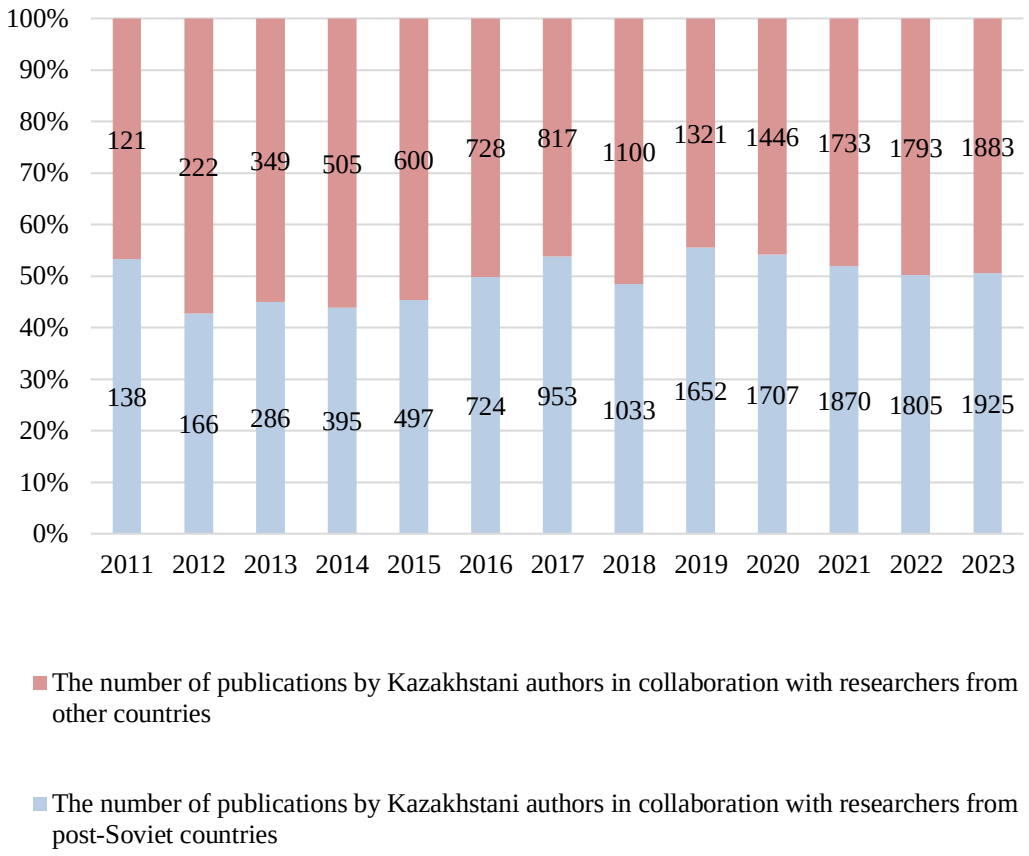
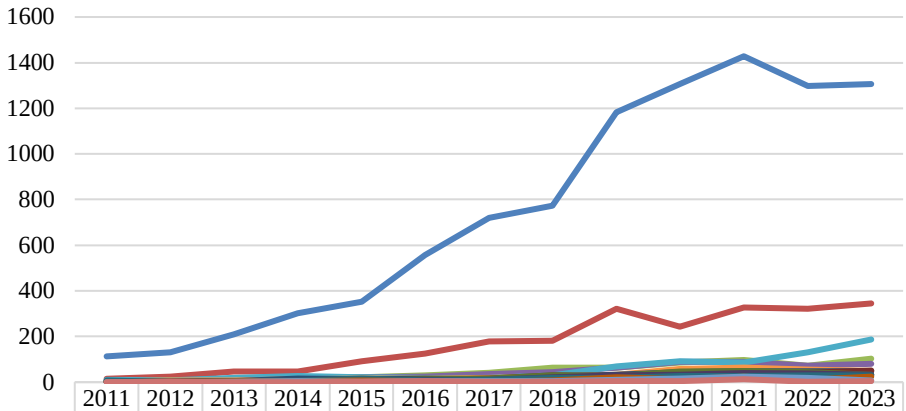


Figure 3. The number and share of publications by Kazakhstani authors in collaboration with researchers from post-Soviet countries and other countries.

This suggests that Kazakhstani authors continue to leverage existing ties or establish new connections with researchers from former Soviet Union countries. Notably, many of these countries are also pursuing policies to internationalize their science and higher education sectors by increasing the number of publications in international journals (Figure 4).



	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Russia	113	132	210	301	352	557	719	774	1182	1307	1428	1298	1306
Ukraine	15	23	48	47	92	124	178	181	321	244	326	322	345
Kyrgyzstan	5	8	12	20	22	31	41	65	65	85	98	71	103
Belarus	7	5	15	14	21	23	36	43	63	89	89	73	81
Uzbekistan	10	8	20	26	21	19	20	30	70	91	87	131	187
Latvia	3	8	8	14	16	17	20	25	32	59	64	55	49
Lithuania	2	3	6	16	13	16	14	24	32	43	48	48	48
Azerbaijan	1	2	3	6	9	10	15	29	28	31	33	41	51
Estonia	3	1	1	6	5	13	15	31	28	46	49	45	34
Georgia	2	3	2	7	7	9	11	21	29	32	40	38	34
Armenia	7	1	6	8	15	7	11	18	19	27	27	29	30
Tajikistan	1	1	5	6	7	4	3	7	22	12	24	9	23
Moldova	0	0	2	4	3	6	5	6	11	16	24	18	8
Turkmenistan	0	1	0	0	1	5	1	2	4	5	13	3	6

Figure 4. The number of publications by Kazakhstani authors with scientists from post-Soviet countries in Scopus journals, 2011-2023.

The decrease in the number of publications by Kazakhstani scientists in collaboration with Russian authors may be related to a particular factor. In 2022, following the imposition of sanctions by Western countries against Russia, the authorized body in Russia's science sector introduced a moratorium on the recognition of articles published in Scopus/WoS journals as part of the qualification evaluation for scientific and teaching personnel (Vedomosti, 2022).

On the other hand, there is an increase in the number of joint publications between Kazakhstani scientists and authors from certain post-Soviet countries. This trend is primarily influenced by the adoption of policies in these countries aimed at increasing the visibility of their research results in the global research community (Berekeyeva et al., 2024). For example, in 2018, Uzbekistan adopted an evaluation system for the qualifications of candidates for academic degrees and titles based on publications in WoS and Scopus journals. As a result, the number of publications by Kazakhstani authors in collaboration with Uzbek colleagues more than doubled in 2019 compared to 2018. Accordingly, authors from post-Soviet countries may be

seeking collaborators from other former Soviet countries to publish joint articles. The shared historical background makes such collaborations more accessible than establishing new ties with scholars from other countries.

Given that Russia leads as the country with which Kazakhstani authors have the highest number of publications, Russian institutions are also represented in the top 10 international collaborators. Specifically, seven foreign organizations, with which Kazakhstani authors have published the most articles in Scopus-indexed journals from 2013 to 2022, are Russian, while one institution each is from Poland, France, and Ukraine (Figure 5).

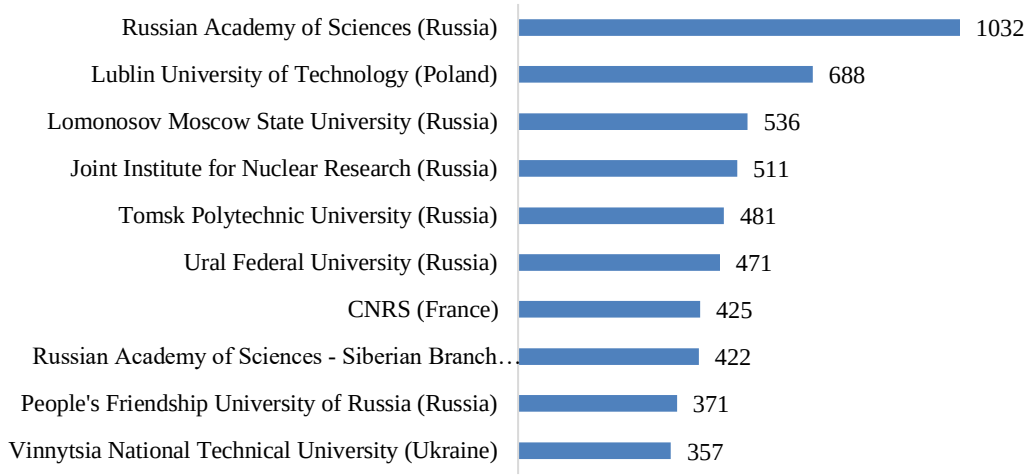


Figure 5. Top 10 foreign institutions with which Kazakhstani authors have published the most articles in Scopus-indexed journals, 2013-2022.

According to SciVal data, between 2014 and 2023, the share of publications by Kazakhstani authors written in international collaboration is 51.2%. Among the 73 considered Kazakhstani organizations, only the following groups have values exceeding this figure (51.2%): universities where teaching is predominantly conducted in English and/or another foreign language (Nazarbayev University – NU, Kazakh-British Technical University – KBTU, KIMEP University, Suleyman Demirel University – SDU) – 62.4%; research institutes (RI) – 56.5%; and organizations based in the capital (Astana) – 55.9% (Table 2).

Table 2. Comparative analysis of international collaboration indicators according to SciVal data for 73 Kazakhstani organizations, grouped into categories, from 2014 to 2023.

	Number of publications with post-Soviet authors	*Number of publications in international collaboration*	*Share of publications with post-Soviet authors*	*Share of publications in international collaboration*

Total across 73 considered organizations	14 665	28 348	51,73%	48,23%
Universities	12 337	24 963	49,42%	47,41%
RI	1 797	2 667	67,38%	56,54%
Organizations in Astana	3 608	9 732	37,07%	55,98%
Organizations in Almaty	6 648	12 612	52,71%	47,51%
Regional organizations	4 409	6 004	73,43%	40,45%
Organizations in Astana and Almaty	10 256	22 344	45,90%	50,86%
Excluding NU, KIMEP, KBTU, SDU	13 652	21 975	62,13%	45,25%
Only NU, KIMEP, KBTU, SDU	1 013	6 373	15,90%	62,42%

The median share of publications by Kazakhstani authors from the 73 organizations considered over the period 2014-2023, co-authored with researchers from post-Soviet countries in relation to the total number of publications in international collaboration, stands at 64.47%. This value is exceeded by the following groups: regional organizations (73.4%) and research institutes (67.4%). Only 22 (30.1%) of the 73 Kazakhstani organizations have a share of publications with post-Soviet authors that is less than 50% of their total publications in international collaboration (Table 3).

Table 3. Indicators of international collaboration according to SciVal for 73 Kazakhstani organizations over the period 2014-2023.

Institutions	*Number of publications with post-Soviet authors*	*Number of publications in international collaboration*	*Share of publications with post-Soviet authors*	*Organization type*	*City*
Ministry of Energy of the Republic of Kazakhstan	136	142	95,77%	Ministry	Astana
Yessenov University	137	147	93,20%	University	Aktau
Institute of Nuclear Physics, National Nuclear Center of the	794	863	92,00%	RI	Almaty

Republic of Kazakhstan					
Karaganda Economic University of Kazpotrebsoyuz	74	81	91,36%	University	Karagandy
Karaganda State Technical University	402	459	87,58%	University	Karagandy
Rudny Industrial Institute	99	116	85,34%	University	Rudny
Pavlodar State Pedagogical University	61	72	84,72%	University	Pavlodar
Zhangir Khan West Kazakhstan Agrarian - Technical University	148	177	83,62%	University	Uralsk
Caspian University	35	42	83,33%	University	Almaty
KAZGUU University	33	40	82,50%	University	Astana
S. Toraighyrov Pavlodar State University	266	323	82,35%	University	Pavlodar
Korkyt Ata Kyzylorda State University	137	169	81,07%	University	Kyzylorda
National Nuclear Center of the Republic of Kazakhstan	132	163	80,98%	RI	Kurchatov
Karaganda State Industrial University	105	132	79,55%	University	Karagandy
South Kazakhstan Medical Academy	83	108	76,85%	University	Shymkent
M.Kh. Dulaty Taraz State University	196	258	75,97%	University	Taraz
Kazakh Research Institute of Processing and Food Industry	50	66	75,76%	RI	Almaty
South Kazakhstan State University (SKSU)	325	429	75,76%	University	Shymkent
Sarsen Amanzholov East Kazakhstan State University	103	137	75,18%	University	Ust-Kamenogorsk
Atyrau Oil and Gas University	36	48	75,00%	University	Atyrau

M. Kozybayev North Kazakhstan State University	84	112	75,00%	University	Petropavlovsk
Academy of Logistics and Transport	131	175	74,86%	University	Almaty
Shakarim University	175	234	74,79%	University	Semey
Esil University	44	59	74,58%	University	Astana
Buketov Karaganda State University	357	479	74,53%	University	Karagandy
D. Serikbayev East Kazakhstan Technical University	389	523	74,38%	University	Ust-Kamenogorsk
Baitursynov Kostanay Regional University	132	181	72,93%	University	Kostanay
Dosmukhamedov Atyrau University	62	86	72,09%	University	Atyrau
Institute of Ionosphere	47	66	71,21%	RI	Almaty
Ministry of Education and Science of the Republic of Kazakhstan	294	417	70,50%	Ministry	Astana
Institute of Information and Computational Technologies	296	422	70,14%	RI	Almaty
L.N. Gumilyov Eurasian National University	1781	2547	69,93%	University	Astana
Almaty Technological University	178	256	69,53%	University	Almaty
K. Zhubanov Aktobe Regional State University	138	206	66,99%	University	Aktobe
National Academy of Science of Kazakhstan - NASK	61	92	66,30%	Academy of Sciences	Almaty
Abay Kazakh National Pedagogical University	300	461	65,08%	University	Almaty
Turan University	49	76	64,47%	University	Almaty
Khoja Akhmet Yassawi International	259	405	63,95%	University	Turkestan

Kazakh-Turkish University					
Almaty Institute of Power Engineering and Telecommunication	259	410	63,17%	University	Almaty
Ualikhanov Kokshetau State University	104	167	62,28%	University	Kokshetau
Zhetysu State University named after I. Zhansugurov	33	53	62,26%	University	Taldykorgan
Kazakh National Women's Teacher Training University	66	109	60,55%	University	Almaty
Ministry of Health of the Republic of Kazakhstan	40	67	59,70%	Ministry	Astana
Astana IT University	85	146	58,22%	University	Astana
Institute of Mathematics and Mathematical Modelling	238	417	57,07%	RI	Almaty
Astana Medical University	167	296	56,42%	University	Astana
West Kazakhstan Marat Ospanov State Medical University	140	249	56,22%	University	Aktobe
Kazakh National Medical University	314	559	56,17%	University	Almaty
Satbayev University	916	1658	55,25%	University	Almaty
Kazakh Ablai Khan University of International Relations and World Languages	16	29	55,17%	University	Almaty
Saken Seifullin Kazakh Agrotechnical University	275	510	53,92%	University	Astana
Institute of Combustion Problems	76	158	48,10%	RI	Almaty
Karaganda State Medical Academy	89	186	47,85%	University	Karagandy

South Kazakhstan State Pedagogical University	28	49	45,90%	University	Shymkent
Almaty Management University	38	83	45,78%	University	Almaty
International Educational Corporation	37	81	45,68%	University	Almaty
NASK - Zoology Institute of the Republic of Kazakhstan	34	75	45,33%	RI	Almaty
Al Farabi Kazakh National University	1860	4110	45,26%	University	Almaty
Semey Medical University	115	255	45,10%	University	Semey
Narxoz University	50	111	45,05%	University	Almaty
National Center for Biotechnology	64	149	42,95%	RI	Astana
Kazakh National Agrarian University	295	693	42,57%	University	Almaty
University of International Business	18	43	41,86%	University	Almaty
Kazakh-British Technical University	266	675	39,41%	University	Almaty
International Information Technology University	89	226	39,38%	University	Almaty
Academy of Public Administration under the President of the Republic of Kazakhstan	11	29	37,93%	University	Astana
Institute of Plant Biology and Biotechnology, Almaty	39	118	33,05%	RI	Almaty
Suleyman Demirel University, Almaty	54	197	27,41%	University	Almaty
Research and Production Center of Microbiology and Virology	12	49	24,49%	RI	Almaty
Institute of Geological Sciences Kazakhstan	12	60	20,00%	RI	Almaty

Nazarbayev University	678	5330	12,72%	University	Astana
KIMEP University	15	171	8,77%	University	Almaty
Kazakhstan Highway Research Institute	3	61	4,92%	RI	Almaty
	14 665	28348	**51,73%**		

The internationalization of science is least developed in the regions of Kazakhstan (i.e., excluding the cities of Almaty and Astana). Thus, the share of publications in international collaboration in relation to the total number of publications in regional organizations is 40.45%, with nearly 3/4 of the articles co-authored with foreign researchers being from post-Soviet countries. This indicates that scientists from the regions of Kazakhstan predominantly collaborate with authors from former Soviet Union countries, whereas connections with researchers from other countries may not be as well developed. For example, at Yessenov University and Karaganda Economic University of Kazpotrebsoyuz, more than 90% of the articles in international collaboration are co-authored with authors from the post-Soviet space (Table 3).

Moreover, stronger scientific ties with post-Soviet researchers are observed among research institute staff: 2/3 of publications in international collaboration are written jointly with authors from the former Soviet Union. This trend is characteristic of those research institutes that were involved in classified developments during the Soviet era. For example, the share of articles with post-Soviet authors in relation to the total number of publications in international collaboration at the Institute of Nuclear Physics, National Nuclear Center of the Republic of Kazakhstan, and National Nuclear Center of the Republic of Kazakhstan is 92% and 81%, respectively (Table 3).

Large international collaborations, defined as publications co-authored by more than 1,000 authors, account for only 7 articles, or 0.17% of all internationally co-authored publications from 2014 to 2023. In contrast, articles co-authored by more than 100 researchers amount to 33 publications (0.8%), while those with more than 50 authors total 47 publications (1.14%).

According to data from Scopus/SciVal, 56% of the articles by Kazakhstani authors in the social sciences, out of all internationally co-authored publications from 2013 to 2022, were written in collaboration with researchers from post-Soviet countries. This percentage increased from 21.4% in 2013 to 56.5% in 2022. Similarly, in the humanities and arts, Kazakhstani authors published 66.2% of their internationally co-authored articles with researchers from post-Soviet countries during the same period.

Thus, the analysis indicates that Kazakhstani researchers in the social sciences, as well as in the humanities and arts, are more likely to collaborate with authors from post-Soviet countries. Several factors may explain this trend. First, national science policies in post-Soviet states and institutional requirements, often aligned with Scopus metrics, serve as a unifying factor, encouraging researchers from the former USSR to collaborate and publish in Scopus-indexed journals to meet national and

institutional criteria for scientific productivity. Second, post-Soviet authors share a Soviet-era approach to academic writing (Yessirkepov et al., 2015), which remains particularly evident in the social sciences and humanities. This is compounded by lower English proficiency (Demeter, 2019), limited exposure to diverse research methodologies, and insufficient familiarity with the publishing standards of English-language journals (Kurambayev & Freedman, 2021). Consequently, some post-Soviet, including Kazakhstani, researchers co-author papers that are frequently published in journals of questionable reputation. The methodological legacy of the Soviet academic tradition in these fields does not align with the expectations of contemporary English-language journals indexed in Scopus. Furthermore, many Scopus-indexed journals require diverse methodologies that are underdeveloped or less widely recognized in post-Soviet countries, contributing to lower research quality. Additionally, scholars in the social sciences and humanities often struggle with adapting to the article structure and formatting standards required by international journals. In contrast, researchers in the natural sciences are generally more familiar with international publishing standards due to the universal nature of scientific methods and their active participation in global research projects.

An analysis of international collaboration indicators among Kazakhstani authors using the SciVal analytics tool showed that articles co-authored with scholars from post-Soviet countries are, on average, cited more frequently than those written with researchers from other countries (Table 7).

Table 7. Comparison of Kazakhstani authors international collaboration metrics according to SciVal, Elsevier, 2014-2023.

	Citation Count	*Citations per Publication*	*Q1**	*Q2*	*Q3*	*Q4*	*Scholarly Output*	*Share of Citations in International Collaboration*	*Share of Scholarly Output in International Collaboration*
Collaboration with post-Soviet countries	249 001	19,87	22,45%	19,96 %	29,83 %	27,76%	12 533	60,44%	50,98%
Collaboration with other countries	162 972	13,52	46,73%	24,12 %	18,71 %	10,44%	12 053	39,56%	49,02%

* - Journal quartiles by Scopus CiteScore Percentile

The average Citations per Publication of Kazakhstani authors' publications co-authored with post-Soviet colleagues between 2014-2023 is 19.87, whereas publications with scholars from other countries equals to 13.52. This is despite the fact that both groups account for approximately half of all articles written in international collaboration. Consequently, publications with post-Soviet scholars were cited 249,001 times (60.44% of all citations for Kazakhstani authors' articles in

international collaboration), while those with authors from other countries were cited 162,972 times (39.56%).

It is worth noting, however, that a large proportion of articles co-authored with researchers from post-Soviet countries were published in journals ranked in the third (Q3) and fourth (Q4) quartiles based on CiteScore percentile: 29.83% and 27.76%, respectively. In contrast, publications with scholars from other countries were published in journals from quartiles 1 and 2 (Q1 and Q2) – 46.73% and 24.12%, respectively.

Thus, the effectiveness (Citations per Publication) of Kazakhstani authors' publication activity with scholars from post-Soviet countries, according to Scopus, is higher than with authors from other countries. This success may be attributed to a shared linguistic environment, which facilitates communication between authors, or to other factors, such as long-term involvement in research projects on a specific topic, which fosters stable scientific groups and, consequently, higher productivity. Another factor might be that a certain group of scholars, particularly those who worked on classified research during the Soviet era, harbor stereotypes of mistrust toward foreign researchers who might steal their ideas. In contrast, there is a degree of mutual trust among post-Soviet researchers.

Conclusion

Thirty-three years after the collapse of the Soviet Union, the scientific connections of Kazakhstani researchers with other post-Soviet authors, as demonstrated by the conducted analysis, have not only remained but even strengthened. Paradoxically, this trend was influenced by the implementation of an indicator-based research evaluation system and policies for the internationalization of science and higher education adopted in post-Soviet countries, including Kazakhstan. These changes prompted researchers to collaborate with colleagues from other countries, with former Soviet Union countries being the first to establish scientific ties. In Kazakhstan, nearly 70% of organizations have more than half of their international collaborations with post-Soviet countries.

Moreover, the effectiveness of these collaborations (with post-Soviet authors) has shown positive results within the Kazakhstani context, with higher average citation rates for articles compared to those co-authored with researchers from other countries.

For a more detailed assessment of the effectiveness of international collaboration and a comparison of the publication activity of Kazakhstani authors with post-Soviet scholars and those from other countries, an analysis of Kazakhstan's emerging institutions is required. For instance, universities where teaching is predominantly conducted in English and/or another foreign language. This analysis will allow for a determination of the level of international collaboration within these institutions, the quality of journals in which their employees publish in collaboration with foreign colleagues, and the average citation rates of these articles. This will provide a more objective comparison of the effectiveness of collaborations between Kazakhstani new formation institutions (which mostly collaborate with non-post-Soviet

countries) and organizations where post-Soviet scientists are the predominant collaborators.

A limitation of this study is that, in certain cases, Kazakhstani authors' joint articles with post-Soviet scholars may have been part of a larger international collaboration that also included representatives from other countries, not solely former Soviet Union states.

Acknowledgments

This work was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan [Grant No. AP19678110].

References

Abalkina, A. (2023). Publication and collaboration anomalies in academic papers originating from a paper mill: Evidence from a Russia-based paper mill. *Learned Publishing*, *36*(4), 689–702. https://doi.org/10.1002/leap.1574

Berekeyeva, A., Sharplin, E., Courtney, M., & Sagitova, R. (2024). Ethical human participant research in Central Asia: a quantitative analysis of attitudes and practices among social science researchers based in the region. *Research Ethics*, *20*(2), 304-330. https://doi.org/10.1177/17470161231202232

Chankseliani, M. (2017). Charting the development of knowledge on Soviet and post-Soviet education through the pages of comparative and international education journals. *Comparative Education*, *53*(2), 265–283. https://doi.org/10.1080/03050068.2017.1293407

Chankseliani, M., Lovakov, A. & Pislyakov, V. (2021). A big picture: bibliometric study of academic publications from post-Soviet countries. *Scientometrics*, *126*, 8701–8730. https://doi.org/10.1007/s11192-021-04124-5

Demeter, M. (2019). The winner takes it all: International inequality in communication and media studies today. *Journalism & Mass Communication Quarterly*, *96*(1), 37–59.

Kerimkulova, S., & Kuzhabekova, A. (2017). Quality assurance in higher education of Kazakhstan: A review of the system and issues. *The rise of quality assurance in Asian higher education*. https://doi.org/10.1016/B978-0-08-100553-8.00006-9

Kudaibergenova, R., Uzakbay, S., Makanova, A., Ramadinkyzy, K., Kistaubayev, E., Dussekeev, R. & Smagulov, 613 K. (2022). Managing publication change at Al-Farabi Kazakh National University: A case study. 614 *Scientometrics*, *127*(1), 453–479. https://doi.org/10.1007/s11192-021-04139-y

Kurambayev, B., & Freedman, E. (2021). Publish or perish? The steep, steep path for Central Asia journalism and mass communication faculty. *Journalism & Mass Communication Educator*, *76*(2), 228–240. https://doi.org/10.1177/1077695820947259

Marina, T. & Sterligov, I. (2021). Prevalence of potentially predatory publishing in Scopus on the country level. *Scientometrics, 126*, 5019–5077. https://doi.org/10.1007/s11192-021-03899-x

Matveeva, N., Batagelj, V., & Ferligoj, A. (2023). Scientific collaboration of post-Soviet countries: The effects of different network normalizations. *Scientometrics*, *128*(8), 4219–4242. https://doi.org/10.1007/s11192-023-04752-z

MES (Ministry of Education and Science of Kazakhstan) (2011a). Law of the Republic of Kazakhstan dated February 18, 2011 No. 407-IV "On Science." Available at https://online.zakon.kz/document/?doc_id=30938581#pos=57

MES (2011b). Order of the Minister of Education and Science of the Republic of Kazakhstan dated March 31, 2011 No. 127 "On approval of the Rules for awarding degrees." Available at https://adilet.zan.kz/rus/archive/docs/V1100006951/31.03.2011

MES (2011c). Order of the Minister of Education and Science of the Republic of Kazakhstan dated March 31, 2011 No. 127 "On approval of the Rules for conferring academic titles (associated professor (docent), professor)." Available at https://adilet.zan.kz/rus/archive/docs/V1100006939/31.03.2011

Moldashev, K., Arystanbaeva, S., & Tleuov, A. (2019). Politika internatsionalizatsii issledovaniy v Kazakhstane: reaktsiya uchenykh [Research internationalization policy in Kazakhstan: Reaction of scholars]. *Central Asian Economic Review*, *127*(4). 85-95. (In Russ.)

Moldashev, K., Arystanbayeva, S., Kozhakhmet, S., & Tleuov, A. (2020). *Problemy integratsii kazakhstanskikh uchenykh v global'nom nauchnom prostranstve* [Problems of integration of Kazakhstani scientists into the global scientific space]. Almaty, Narxoz University. (In Russ.)

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of web of science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–218. https://doi.org/10.1007/s11192-015-1765-5

Smagulov, K., Makanova, A., & Burshukova, G. (2018). Analiz naukometricheskikh pokazateley publikatsionnoy aktivnosti kazakhstanskikh avtorov v izdaniyakh, vkhodyashchikh v bazu dannykh Scopus [Analysis of scientometric indicators of Kazakhstani authors' publication activity in journals, included in the Scopus database]. *Vestnik KazNU. Seriya Ekonomicheskaya*, *123*(1), 233–253. (In Russ.)

Turginbayeva, A., & Makanova, A. (2024). Otsenka kvalifikatsii uchenykh na osnove publikatsiy v nauchnykh izdaniyakh baz dannykh Scopus i Web of Science v normativno-pravovykh aktakh Kazakhstana [Assessment of the qualifications of scientists based on publications in scientific journals of the Scopus and Web of Science databases in regulatory documents of Kazakhstan]. *Vestnik KazNU. Seriya Yuridicheskaya*, *110*(2), 12–30. https://doi.org/10.26577/JAPJ2024-110-b-02 (In Russ.)

Vedomosti (2022). V Rossii vveden moratoriy na pokazateli po publikatsiyam uchenykh v zarubezhnykh zhurnalakh [A moratorium has been introduced in Russia on evaluating researchers based on their publications in international journals]. Available at https://www.vedomosti.ru/society/news/2022/03/21/914514-moratorii-na-pokazateli-po-publikatsiyam (In Russ.)

Yessirkepov, M., Nurmashev, B., & Anartayeva, M. (2015). A Scopus-based analysis of publication activity in Kazakhstan from 2010 to 2015: Positive trends, concerns, and possible solutions. *Journal of Korean Medical Science*, *30*(12), 1915–1919. https://doi.org/10.3346/jkms.2015.30.12.1915

https://doi.org/10.51408/issi2025_025

Knowledge Combination and Research Impact: A Comparison of Sources and Keywords Co-Citation

Hsiao Tsung-Ming[1], Tang Muh-Chyun[2]

[1]*162228@mail.tku.edu.tw*
Department of Information and Library Science, Tamkang University (Taiwan)

[2]*mctang@ntu.edu.tw*
Department of Library and Information Science, National Taiwan University (Taiwan)

Abstract

Combining knowledge from diverse origins has long been recognized as a key driver of innovation. Although many studies have examined how such combinations influence research impact, their findings remain inconsistent. One potential reason is the use of different units of measurement for novelty and conventionality. Using data from the DBLP Citation Network, this study compares two approaches—one based on sources and the other on keywords co-cited frequency—to measure research novelty. We found a low correlation between these two measures, suggesting that each captures distinct aspects of novelty. In line with Uzzi et al. (2013), it was found that papers exhibiting both high novelty and high conventionality (HNHC) are more likely to achieve high citation impact, especially when novelty is measured at the source level. Logit regression indicates that source-based HNHC is a strong predictor of highly-cited "hit" papers, though the keyword-based measure also contributes a smaller but statistically significant effect. These results highlight the importance of carefully selecting units of analysis when investigating the relationships between novelty, conventionality, and research impact.

Introduction

The synthesis of heterogeneous knowledge has long been recognized as a key driver of innovation. However, recent studies suggest that achieving high-impact research often requires balancing high novelty with strong conventionality (Uzzi et al., 2013). This is an intriguing development as previous studies of the relationships between novelty and research impact had often overlooked the need to situate novelty within conventional wisdom. Methodologically, research novelty is frequently measured by the rarity or unexpectedness of knowledge combined in a paper. A paper is considered novel if it synthesizes knowledge units that appear for the first time or occur rarely. Two types of knowledge units have been proposed to measure novelty: one based on the journals cited and the other on the keywords or subject headings used to index the paper. However, little research has examined the consistency of the novelty assessments produced by these two approaches. To address this gap, the present study compared novelty and conventionality measurements derived from source co-citation (journals) and keyword co-citation using DBLP, a large citation network dataset in the field of computer science. Additionally, the study evaluated how effectively combinations of novelty and conventionality, as measured by each approach, can identify highly cited papers. citation. A novel aspect of this research is the use of keyword co-citation, rather than keyword co-occurrence, as an indicator of a paper's novelty. Our findings reveal a slight correlation between journal-based

and keyword-based co-citation measures of novelty, suggesting they are capturing different aspects of novelty. Interestingly, the combination of high novelty and conventionality was associated with greater odds of producing a hit paper when cited journal is used as the basic knowledge unit.

Literature Review

Creating innovative ideas relies on combining knowledge from various sources, this is especially so when the combination is novel. In the last decades, various quantitative indicators have been proposed to measure novelty based on how rare and different the combination of knowledge unit is (e.g. Bornmann et al., 2019; Carayol et al., 2019;). While investigating the relationship between exploring new possibilities and exploiting established certainties in organizational learning, March (1991) found that the process of innovation relates to refining the existing technical combinations and creating new technical combinations. The combinatorial conjecture, "all creativity results from combinations of mental representations", was also evaluated by Thagard (2012) with great scientific discoveries and technological inventions. After surveying 200 invention examples, he supported this conjecture. As claimed by Kaplan and Vakili (2015), novel ideas require both a broad search for information and a process of recombining diverse knowledge. The recombination process plays a critical role in enhancing novelty and producing breakthrough-class papers (Bornmann et al., 2019).

Atypical and Conventional Combination

While novelty is often considered a necessary condition for innovation, it has been pointed out that novelty alone is not enough to drive impact. Uzzi and his colleagues (2013) argued that "balancing atypical knowledge with conventional knowledge may be critical to the link between innovativeness and impact" (p. 468). To determine the degree of how atypical or conventional an article's knowledge combinations are, they examined the sources of knowledge, namely the journals listed in its bibliography. Specifically, they built the journal co-citation networks by year with Web of Science (WoS) data and aggregated the frequency of journal pairing, two journals co-cited by articles. The atypical or conventional level of a combination was determined by comparing its observed frequency with the expected frequency, derived from a randomized simulation network that retains key features of its corresponding journal co-citation network. A journal pair was classified as an atypical combination if its observed frequency lowers than the expected frequency. Conversely, if a journal pair occurs frequently than expected, it is considered as conventional combination. The observed frequency of the journal pair was converted to a z-score to facilitate comparison.

For each paper, two summary statistics, novel and conventional values, were derived from the rank-ordering of the z-scores of all its journal pairings. As depicted in Figure 1, the novelty, left tail, was defined as the 10th percentile z-score, and the conventionality was defined as the median z-score. Novelty serves as a criterion for classifying papers into high or low novelty, and conventionality can be applied in a similar manner. Hence, papers can be categorized into one of four quadrants based

on their conventionality and novelty. Uzzi and his colleagues (2013) analyzed 17.9 million articles and 302 million articles references across all WoS disciplines from 1950 to 2000 and showed that articles properly balancing high levels of both novelty and conventionality have the highest potential of becoming high-impact publications. Based on the results, they argued that effectively embedding novel ideas into established traditions is the key drivers of scientific advancement.

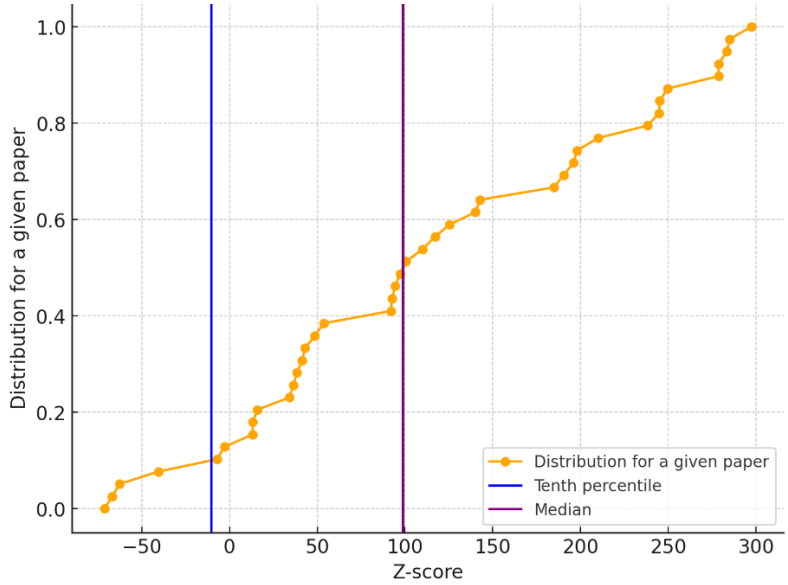


Figure 1. The distribution of z-scores of an article's journal pairings and how the novelty and conventionality of this article are determined. Redraw based on Uzzi et al. (2013).

Source-based Approach

The method proposed by Uzzi et al. (2013) is based on how the sources of references are co-cited by papers. This source-based approach is applied by several studies, and one of them is Boyack & Klavans (2014). They used Scopus data of articles published from 2001 to 2010 to replicate the findings of Uzzi et al. (2013) and further explore the disciplinary effects on the relations between paper's atypical and conventional combinations and its probability of being highly-cited papers. Instead of deciding the expected frequency of the source combination with simulated citation network, K50 was used to determine the novel/conventional degrees of a journal pair was determined by K50, a method examined by Klavans & Boyack (2004). While affirming the findings of Uzzi and his colleagues, Boyack & Klavans (2014) highlighted the potential mediating effects of disciplines and publication venues. According to their findings, the relationship was less evident when identifying the top 5 percent of highly-cited papers within individual disciplines, compared to findings across all disciplines (Uzzi et al., 2013). Further investigation of the top 20 highly-cited journals revealed that leading physics journal typically exhibited high conventionality and low novelty. Conversely, top biomedical journals combined

high novelty with high conventionality. Meanwhile, multidisciplinary journals like *Nature* and *Science* exhibited high novelty but low conventionality.

Source-based approach is used by two later studies, Lee et al. (2015) and Wang et al. (2017). Lee et al. (2015) analyzed 9,428 WoS-indexed publications, covering publication years 2001 to 2006, to examine how team size, field diversity, and task diversity influence creativity. Building on Uzzi et al. (2013), they defined the commonness of a journal pairing in a specific year as the fraction of its observed frequency to its expected frequency. In their research, the expected frequency is calculated as the total number of all journal pairings multiplied by the joint probability of the co-occurrence of the two journals. Instead of using co-occurrence of the two journals directly, Wang et al. (2017) proposed that the similarity between two journals can be defined as the cosine similarity of their corresponding row vectors, extracted from the journal co-citation matrix, as shown in Figure 2. The higher the similarity of a journal pairing, the lower its novelty. Hence, a paper's novelty is measured by the sum of the differences in all its journal pairings.

	J_1	J_2	J_3	J_4	J_5	…
J_1	/	0	3	0	5	…
J_2	0	/	6	2	3	…
J_3	3	6	/	5	4	…
J_4	0	2	5	/	0	…
J_5	5	3	4	0	/	…
…	…	…	…	…	…	/

Figure 2. The journal co-citation matrix. The number in each cell represents the frequency that two journals are co-cited. Redraw based on Wang et al. (2017).

The source-based approaches proposed by studies reviewed above are examined by Bornmann et al. (2019) and Fontana et al. (2020). Bornmann et al. (2019) used the human recommendations of papers from F100Prime, a post-publication peer review system, to evaluate the validity of two novelty metrics proposed by Lee et al. (2015) and Wang et al. (2017), referred to as novelty score U and novelty score W, respectively. While novelty U followed the unexpected combinations in Uzzi et al. (2013), novelty W counted the number of novel combinations in the references. In addition, they introduced a novelty score K determined by comparing the new keywords to the existing ones in a specific subject category. According to their research findings, novelty score U agreed with their formulated expectations mostly, while novelty score W lacked convergent validity with the FMs' assessments. The logistic regression result revealed that as novelty score K increased, the likelihood of an article being included in F1000 Prime decreased. In another research, Fontana et al. (2020) analyzed the novelty indicators proposed by Uzzi et al. (2013) and Wang et al. (2017) with 230,854 articles published on 8 journals of the American Physical Society. Notably, they used the domain-specific subject classification, instead of journal, as the basic unit based on which the novelty and interdisciplinarity measures were calculated. Infrequent co-occurrences of subject headings were considered

more novel. They showed that novelty score W lacks ability to tell novel and non-novel articles and that novelty score U correlated well with interdisciplinarity.
In summary, a series of studies measure the novelty of a publication based on its references. These studies propose that the novelty level of an article depends on the rarity of its co-citation relationships. The studies reviewed above consider co-citation relationships at the level of sources where the references are published. Instead of focusing on sources, some studies explore how research articles combine topics. The following section will review these related studies.

Topic-based Approach

An alternative approach to defining novelty is by analyzing how articles combine topics. The novelty score K used by Bornmann et al. (2019) is based on comparing the new keywords to existing ones in a specific domain. Besides keywords, some studies use controlled vocabularies like chemical annotations or MeSH to assess novelty. Foster et al. (2015) proposed that "five strategies available to a scientist facing a network of known scientific relationship: *jump*, *new consolidation*, *new bridge*, *repeat consolidation*, or *repeat bridge*" (p. 881). Figure 3 illustrates the five strategies, and these strategies were further divided into two classes: innovation (*jump*, *new consolidation*, and *new bridge*) and tradition (*repeat consolidation* and *repeat*). Traditional strategies involve scientists delving deeper into established knowledge entities and relationships, while innovative strategies involve introducing novel ones. Their study used chemical annotations, extracted from abstracts in the MEDLINE collection, as nodes in the knowledge network, with edges representing the co-occurrence of chemical entities within an abstract. According to their findings, while innovative work has higher impact potential, its rewards do not compensate for the risk of non-publication.

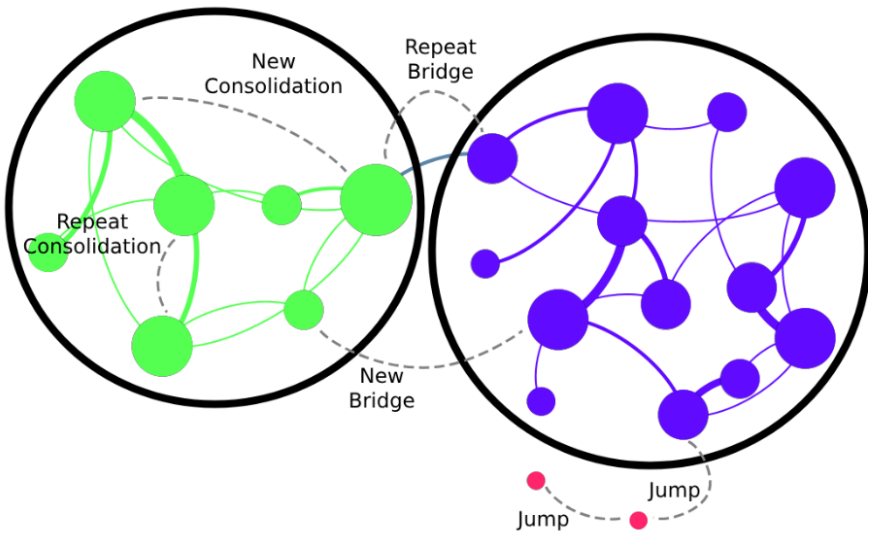


Figure 3. The five strategies in a knowledge network. Bridge connects knowledge entities of two domains, consolidation links knowledge entities with the same domain. Redraw based on Foster et al. (2015).

Instead of using chemical entities, Boudreau et al. (2016) and Ruan et al. (2023) utilized MeSH terms to measure the novelty. To measure novelty, Boudreau et al. (2016) utilized MeSH term combinations and analyzed their appearance in the entire existing related literature. They proposed that novelty was determined by the proportion of term combinations in a given proposal that had not appeared before. The novelty was expressed as a percentile, ranging from 1% (least novel) to 100% (most novel). Ruan et al. (2023) also utilized MeSH term as the unit to determine topic combination novelty. Their design followed Uzzi et al. (2013) and Lee et al. (2015) and adopted "the proportion of the observed and expected frequency of a combination of MeSH terms to denote the *commonness* of the MeSH pair" (p. 5). Carayol et al. (2019) proposed a novelty measure based on the pairwise author keyword co-occurrence in papers indexed in Web of Science in a given year and field. The less common a pair of keywords cooccur, the higher its novelty. It was found that higher novelty was more likely to be observed in larger teams, especially those spanning across institutional and geographic boundaries. Importantly, the correlation between novelty measured through pairwise keyword co-occurrence and journal co-citation was found to be small. Furthermore, pairwise keyword novelty was positively associated with higher citation counts within a three-year citation window, and papers with high novelty had greater odds of becoming "hit papers."

Our review showed increasing efforts in measuring the novelty of a paper, and exploring the relationship between novelty and impact. It is still unclear that, whether novelty along, or the combination of both novelty and conventionality is more conducive to higher impact. Furthermore, as different studies used different knowledge unit for the base of combination, it is difficult to assess how consistent the results are. The potential inconsistence between these using source vs. topic as the base of measuring knowledge combination poses a great challenge to clarify the relationship between knowledge combination and research impact. The purpose of this study is therefore to compare two types of methods, the source-based and topic-based approaches, in measuring an article's novelty and conventionality, and to explore the relationships between the resulting novelty/convention combination and research impact. Specifically, our research questions are as follows:

1. When determining the novelty and conventionality of a given paper, do the source-based and topic-based approaches produce consistent results? This can be further tested by:
a. Are the novel/conventional rank-orders revealed by the two approaches aligned with one another?"
b. When categorizing a paper as high or low in novelty/conventionality, are the classifications from the two approaches consistent?
2. Following Uzzi et al. (2013), do papers integrating high novelty and conventionality (HNHC) resulting in higher odds of being highly cited? And if so,
3. When identifying HNHC, which approach (source vs. keyword-based) yields a higher probability of highly cited papers?
4. Does combining the two approaches offer greater advantage in revealing the relationship between HNHC and high impact?

Research Design

This study uses the DBLP dataset, which comprises over 7 million research articles in computer science. The version, DBLP-Citation-network v13, utilized in this study is maintained by Tang et al. (2008) and has been widely used in various types of research, such as developing recommendation systems (Huang et al., 2024; Kanwal & Amjad, 2024) and predicting scholar impact (Zhang & Wu, 2024). The raw dataset, formatted in JSON, comprises 5,354,306 publications and 48,277,950 citation relations. We extracted publication metadata and citation relations from the dataset. Following the extraction process, publications published prior to 2000 and after 2015 were excluded. Given that both source-based and topic-based approaches were included in this study, any publication with five or fewer references or keywords was excluded to avoid possible bias. After these procedures, 1,725,037 publications were included in this study.

By utilizing the citation relationships in this dataset, the yearly article co-citation networks were built. The source-based approach included in this study was a slightly modified version of the method employed by Boyack and Klavans (2014). Therefore, the article co-citation networks were transferred into source co-citation networks (SCCN) with the procedures reported in the supplementary materials of Uzzi et al. (2013b). The keyword co-citation networks (KCCN) were constructed in similar ways. The source and keywords were the paper venue ID and keywords extracted from the DBLP dataset. Specifically, we used the 'venue.id' field provided in DBLP v13 as the source ID, which indicated the venue in which an article was published. For keywords, we employed the author-provided keywords included in the dataset. The novelty and conventionality for a source pairing or a keyword pairing were determined by K50, a method used for measuring the relatedness of two entities (Boyack & Klavans, 2014; Klavans & Boyack, 2006). For a pair of entities *i* and *j*, their K50 value was calculated using the following formula.

$$K50_{i,j} = K50_{j,i} = max\left[\frac{(F_{i,j} - E_{i,j})}{\sqrt{S_i S_j}}, \frac{(F_{j,i} - E_{j,i})}{\sqrt{S_i S_j}}\right]$$

$$E_{i,j} = S_i S_j / (SS - S_i)$$

$$SS = \sum_{i=1}^{n} S_i$$

$$S_i = \sum_{j=1}^{n} F_{i,j}$$

$F_{i,j}$ denotes the observed frequency with which entities i and j co-occur in the reference documents of a specific year, and $E_{i,j}$ represents the expected frequency. Therefore, any publication included in this study had two distributions of K50, based on its source pairings and keyword pairings. The median of a K50 distribution was the conventionality of a publication. For novelty, we referred to Boyack and Klavans

(2014) and defined the 5th percentile of the K50 distribution as novelty. In addition, alternative thresholds: the 1st and 10th percentile of the K50 distribution were also tested as supplementary novelty metrics to evaluate the robustness of our findings. Each publication in this research features two sets of novelty/conventionality values, calculated separately from SCCN and KCCN.

The publications were classified into four categories based on their novelty/conventionality and their comparison to the novelty/conventionality scores of all publications in the same year. Our study adopts a relative criterion for identifying high/low novelty. High/low novelty was determined by the 40th percentile (PR40) of all novelty scores, and high/low conventionality was determined by the median of all conventionality scores. The four categories are:

- High novelty & high conventionality (HNHC): The novelty value is less than PR40; the conventionality value is higher than the median.
- High novelty & low conventionality (HNLC): The novelty value is less than PR40; the conventionality value lowerthan the median.
- Low novelty & high conventionality (LNHC): The novelty value is higher than PR40; the conventionality value is higher than the median.
- Low novelty & low conventionality (LNLC): The novelty value is higher than PR40; the conventionality value lower than median.

Note that a lower novelty value indicates that the source/keyword pair occurs less frequently, which suggests a novel combination. Therefore, publications with lower novelty values are classified as high novelty.

Results and Discussion

After preprocessing procedures detailed in the research design, a total of 1,725,037 publications were included in this study. The yearly distribution of these publications is presented in Table 1. The number of included publications increases steadily from 32,364 in 2000 to 198,275 in 2015. Tables 2 and 3 provide the statistics for SCCN and KCCN from 2000 to 2015, respectively. The number of sources ranges from 8,372 in 2000 to 26,124 in 2015. Similarly, the number of source pairings grows from 509,928 in 2000 to 4,642,351 in 2015. Overall, the number of sources and source pairings increase three- and ninefold, respectively, during this period. In the same year, the network scale of KCCN is larger than SCCN. Between 2000 and 2015, the number of keywords grows from 45,642 to 101,837, while the number of keyword pairings expands from 21,958,870 to 124,796,940. These figures indicate two- and sixfold increases, respectively.

Table 1. Number of Included Publications.

Year	*Articles*	*Year*	*Articles*	*Year*	*Articles*	*Year*	*Articles*
2000	32,364	2004	63,913	2008	113,347	2012	156,086
2001	35,611	2005	77,090	2009	125,375	2013	172,782
2002	41,209	2006	93,273	2010	134,322	2014	186,702
2003	49,909	2007	102,392	2011	142,387	2015	198,275

Table 2. Yearly statistics of SCCN.

Year	*Source*	*Pairs*	*Year*	*Source*	*Pairs*
2000	8,372	509,928	2008	17,096	1,931,129
2001	9,084	597,745	2009	18,465	2,201,253
2002	9,846	698,292	2010	19,771	2,521,501
2003	10,838	807,826	2011	20,982	2,823,376
2004	11,920	984,009	2012	22,117	3,100,438
2005	13,151	1,232,116	2013	23,277	3,606,916
2006	14,541	1,427,682	2014	24,539	4,022,984
2007	15,671	1,721,678	2015	26,124	4,642,351

Table 3. Yearly statistics of KCCN.

Year	*Keywords*	*Pairs*	*Year*	*Keywords*	*Pairs*
2000	45,642	21,958,870	2008	77,207	63,226,895
2001	48,515	24,176,947	2009	80,913	69,289,002
2002	51,126	27,834,411	2010	84,500	77,334,472
2003	55,244	30,885,836	2011	87.938	83,561,294
2004	59,401	35,953,067	2012	91,750	91,903,280
2005	64,005	43,137,967	2013	95,690	103,920,458
2006	68,679	49,626,908	2014	98,927	112,976,940
2007	72,955	56,844,272	2015	101,837	124,797,226

Rank-Order Similarity and Classification Consistency: Source vs. Keyword Approaches

This study utilized Spearman's rank correlation to investigate whether source-based and keyword-based approaches evaluate the publications' novelty/conventionality consistently. The results are reported in Figure 4. Novelty (1), Novelty (5), and Novelty (10) represent the results based on utilizing the 1st, 5th, and 10th percentile of the K50 distribution as measures of a publication's novelty. The Spearman rank correlation coefficients range from 0.1 to 0.3, suggesting a weak positive correlation. When Novelty (1) is excluded, the coefficients drop below 0.25. The findings suggest that the novelty/conventionality rank orders from the two approaches are weakly related.

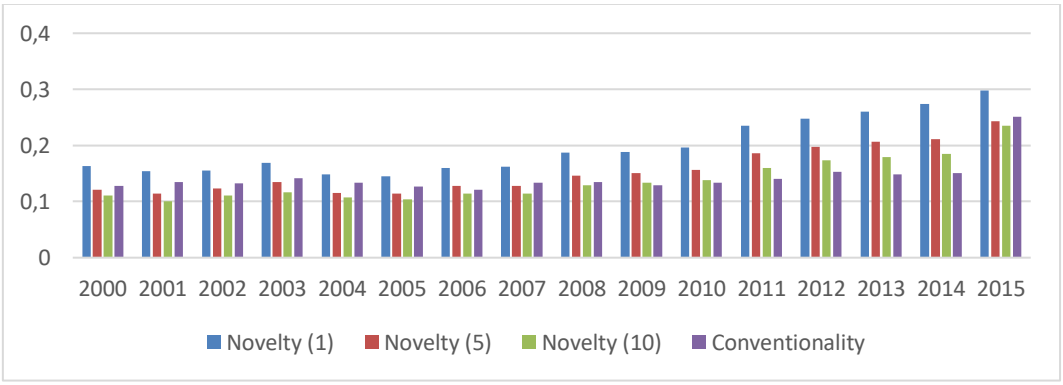


Figure 4. The Spearman rank correlation between two approaches.

We examine whether a publication is classified into the same category by the two approaches. For example, if the source-based approach classifies a publication as high novelty, does the keyword-based approach do the same? Cohen's Kappa, a statistical measure for evaluating agreement between two classifiers, was used to assess the consistency of categorization results between the two approaches. Cohen's Kappa measures the difference between observed agreement and expected agreement by chance. It ranges from -1 to 1, with 1 signifying perfect agreement and 0 indicating agreement equivalent to random chance. Table 4 reported the details. The results indicate that the degree of consistency between the two approaches is only marginally better than what is expected by chance. While consistency has improved over time and increased significantly since 2008, the correlation between the two approaches remains low. We currently suspect that this phenomenon may be attributable to the growth in data size. However, further research is needed to fully address this issue.

Table 4. Classification consistency of binary classes (high/low).

	Novelty (1)	*Novelty (5)*	*Novelty (10)*	*Conventionality*
2000	0.09	0.07	0.06	0.09
2001	0.08	0.07	0.05	0.10
2002	0.09	0.07	0.06	0.09
2003	0.10	0.08	0.06	0.09
2004	0.09	0.07	0.06	0.09
2005	0.09	0.06	0.06	0.08
2006	0.10	0.08	0.08	0.08
2007	0.09	0.08	0.07	0.09
2008	0.12	0.10	0.08	0.09
2009	0.12	0.11	0.09	0.08
2010	0.13	0.11	0.09	0.08
2011	0.16	0.13	0.11	0.09
2012	0.18	0.15	0.12	0.10
2013	0.19	0.16	0.12	0.10
2014	0.19	0.16	0.12	0.10
2015	0.19	0.17	0.16	0.20

Note. Novelty (1) refers to the results of examining the classification consistency of the novelty type derived from two approaches based on the publication's Novelty (1). The same applies to the other notations.

By combining novelty and conventionality values, each approach classifies a publication into one of four possible categories: NHNC, NHLC, LNHC, and LNLC. We further examine the classification consistency of four categories with Cohen's Kappa. Similarly, the degree of consistency between two approaches is weak. Table 5 reports the details. The result indicates that two approaches may evaluate the publication's novelty/conventionality from different perspectives and classify the same publication into various categories.

Table 5. Classification consistency of multiple classes.

	Novelty (1)	*Novelty (5)*	*Novelty (10)*
2000	0.07	0.07	0.06
2001	0.07	0.07	0.06
2002	0.07	0.07	0.06
2003	0.08	0.07	0.06
2004	0.07	0.07	0.06
2005	0.07	0.06	0.06
2006	0.08	0.07	0.07
2007	0.07	0.07	0.06
2008	0.08	0.08	0.07
2009	0.08	0.08	0.07
2010	0.08	0.08	0.07
2011	0.10	0.09	0.08
2012	0.11	0.10	0.09
2013	0.11	0.10	0.09
2014	0.11	0.10	0.09
2015	0.14	0.13	0.13

Note. Novelty (1) refers to the results of examining the classification consistency of the four classes determined by two approaches based on the publication's Novelty (1) and conventionality value. The same applies to the other notations.

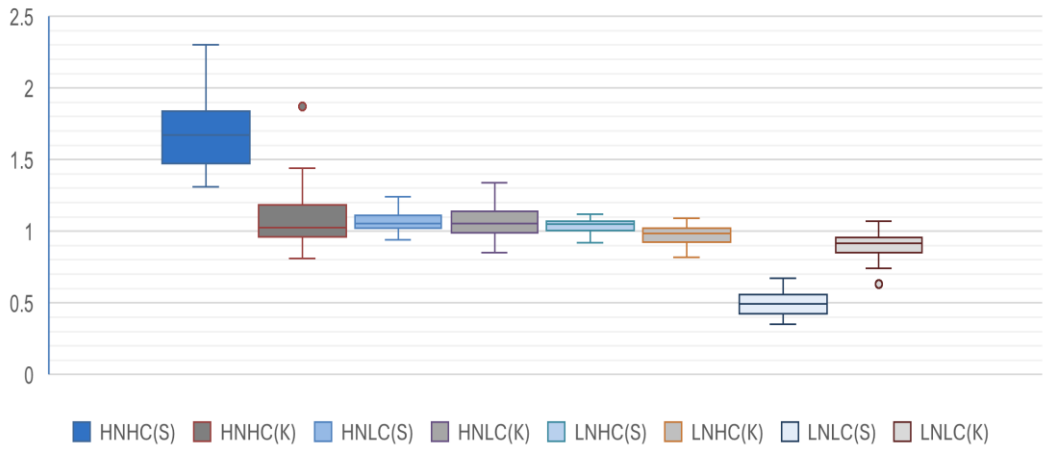
Ability in Identifying Highly Cited Research Publications

To better understand the differences between the two approaches, we analyze the probability that publications categorized into different groups was highly cited. Specifically, we compared the probabilities of being highly cited across the four categories (HNHC, HNLC, LNHC, LNLC) within each approach and investigated whether the probabilities of being highly cited in corresponding categories differ between the source-based and keyword-based approaches. Highly cited publications are defined as those with citation counts in the top 5% for a given year. To ensure the robustness of our findings, we also examine results using alternative thresholds, defining highly cited publications as those in the top 1% and 10%, respectively.

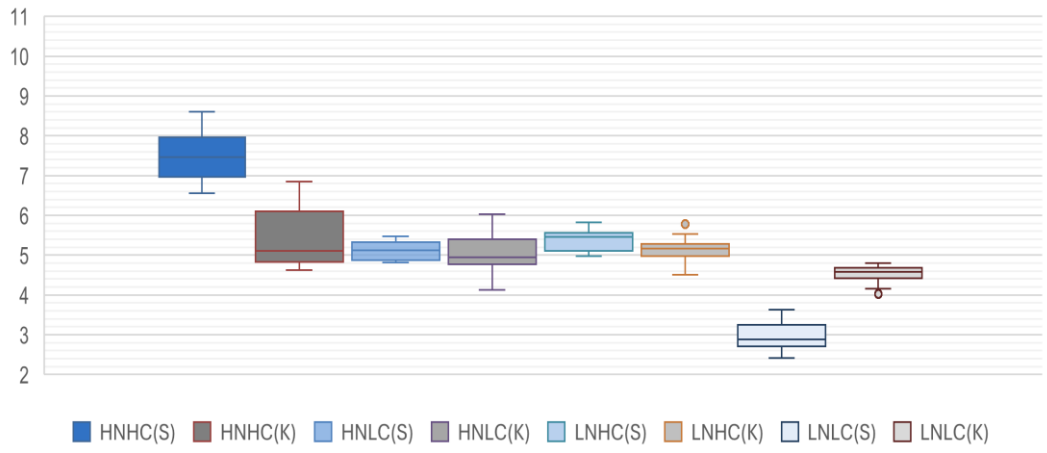
Figure 5 presents the yearly probabilities distribution of being highly cited, with novelty defined as the 5th percentile (PR5) and conventionality as the median of a publication's K50 distribution. We used the notation S and K to denote the combinatory unit of knowledge source and keywords, respectively. Thus, HNHC(S) denotes the category based on the source-based approach, while HNHC(K) denotes the category based on the keyword-based approach. When analyzing the probabilities of being highly cited for groups formed using the source-based approach, the results suggest that in both approaches HNHC has the highest

probability of being highly cited, though the differences is much more salient when journals are used as the basic unit of knowledge.

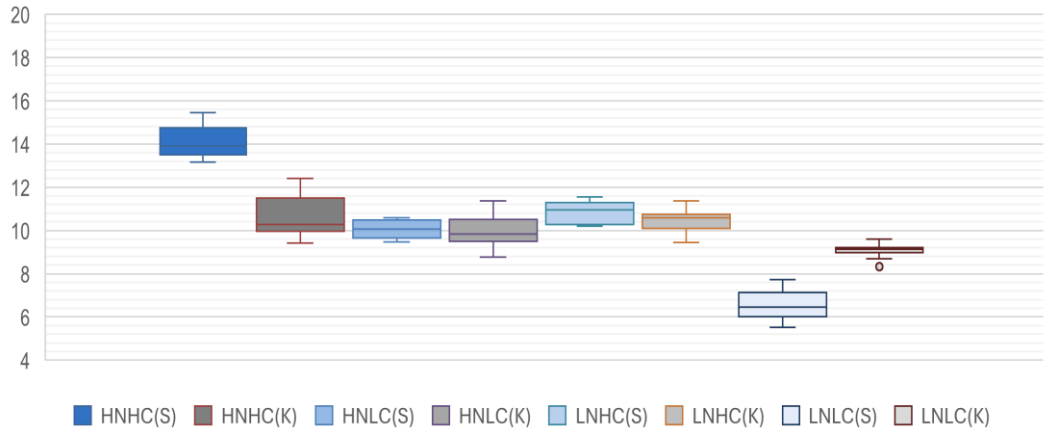
HNHC(S) consistently exhibits nearly double the probability of being highly cited compared to HNLC(S) and LNHC(S). This difference grows to approximately 3–4 times when compared with LNLC(S). However, this pattern is less apparent when examining groups formed using the keyword-based approach. The probabilities of being highly cited for HNHC(K) is only slightly higher than the rest, with only LNLC(K) showing the lowest probability of producing highly cited papers. This pattern remained robust across various thresholds used to define highly cited papers (see Appendices I and II).



(a) Top 1% cited articles as highly-cited articles



(b) Top 5% cited articles as highly-cited articles

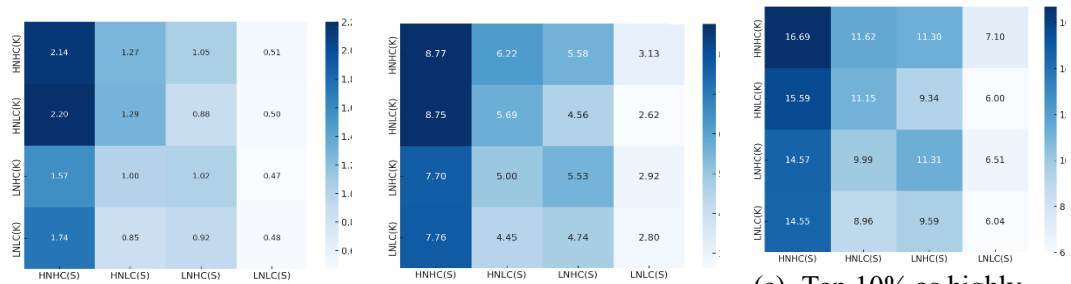


(c) Top 10% cited articles as highly-cited articles

Figure 5. Probabilities of being top highly cited across groups: Novelty (5). Y-axis represents the probability.

Analyzing the Effects of Combining Two Classification Approaches

Given HNHC category yielded the highest probabilities of identifying highly cited papers when either journal or keyword-based approaches was used, though to different extent, journal-based approach is a much stronger predictor. Using a keyword-based approach to measure novelty and conventionality, the difference in the probability of being among the top 5% most-cited articles across the four categories is less than 1% on average. However, while using a journal-based approach, the difference is more the 4%. And as shown earlier, the correlation between these two approaches is low. We explore whether combining both source-based and keyword-based approach enhances the relationship between HNHC and citation impact. In other words, does keyword-based HNHC provide extra explanatory power over and above source-base HNHC is predicting highly cited papers. As shown in Figure 6, if a publication is classified as HNHC by the source-based approach and topic-based approach, its probability of being highly cited is slightly higher.



(a) Top 1% as highly-cited (b) Top 5% as highly-cited (c) Top 10% as highly-cited

Figure 6. Highly-Cited Probabilities for Combined Classifications from Source-Based and Topic-Based Approaches.

To further test our hypothesis, we conducted a logistic regression analysis using highly cited papers as the dependent variable. The source-based and keyword-based HNHC categories served as the two key predictors, allowing us to assess the explanatory power of each classification approach. Specifically, two dummy variables were created to indicate whether an article was categorized as HNHC by the source-based approach and the keyword-based approach, respectively.

As shown in Table 6, As when all predictor variables are set to zero, the model estimates a log-odds of -3.0329, which corresponds to a predicated probability of about 4.6%, closed to our definition for highly-cited articles. Both predictors were significant, despite great differences in their coefficient. If an article is categorized into HNHC by source-based approach, its probability of being highly-cited articles increases to 8.26%, a 79.9% relative increase in the likelihood of the event. Similarly, the probability rises to 5.13%, a smaller 11.6% increase, when an article is categorized into HNHC by topic-based approach. If both approaches classify an article into HNHC, the event probability reaches 9.18%, a nearly 99.8% increase from the baseline.

Table 6. The odds of being highly-cited papers when identified as HNHC by two approaches.

Variable	*Coefficient*	*Stand error*	*Percentage change in odds*
NHNC(S)	0.62***	0.009	79.9
NHNC(K)	0.12***	0.012	11.6
Constant	-3.03***	0.004	

Note. *** <.001

Conclusion

While combination of heterogenous knowledge has long been recognized as a great source for innovation. It remains unclear that novelty along, or the combination of both novelty and conventionality is able to yield high impact research. Novelty has been shown to be associated with frontier research projects (Boudreau et al., 2016), higher research impact (Carayol et al., 2019), and seminal works in scientometrics (Tahamtan & Bornmann, 2018). On the other hand, studies also suggest that novelty alone is not enough, that it is also important to situate the novel ideas in established wisdom, therefore the importance of combining novelty and conventionality (Uzzi et al., 2013; Boyack & Klavans, 2014). One possible explanation for such inconsistency is the use of different knowledge units used when measuring novelty. The first step to clarify the relationship between knowledge combination and impact is therefor to examine how consistent when source based and topic based measurement of the construct of novelty/conventionality. Using DBLP dataset in the domain of computer sciences, we set out to compare how results from topic vs. source based knowledge unit are consistent with each other. The results show that the correlation between the two method is low, suggesting they are capturing different aspect of novelty. Furthermore, Consistent with the original research by Uzzi et al. (2013), we found that a paper combining high novelty and conventionality increases its likelihood of becoming highly cited within a given year. However, this

relationship between high novelty and conventionality and the likelihood of a hit paper was observed was much more salient when novelty and conventionality were calculated using journal co-citation data, and less so when keyword co-citation data was used.

These findings indicate that the source-based approach may better be able highlight the advantages of integrating high novelty and high conventionality, as demonstrated by the increased likelihood of being highly cited. Several limitations need to be noted, one of which is the lack of vocabulary control of author assigned keywords. While it is argued that author keywords, compared by control vocabulary such as MeSH, offers a great granularity therefore more precise representation of the topics (Carayol et al., 2019). Yet, without the benefit of vocabulary control, the actually cooccurrence frequence of topics is likely to highly underestimated because of morphological and semantic variations of the topics. And it is difficult to assess the extent of this underestimation and how this might impact the measurement of novelty. One possible solution to this dilemma is to utilize automatically-assigned topics such as SciVal topics used in Scopus and micro citation topics used in Web of Science. It should also be noted that, instead of keyword co-occurrence is the focal paper, as commonly done in previous research, we have adopted a novel approach of using keywords co-occurrence appearing in the cited references by the focal paper, which resulting a much greater set of keyword co-occurrence pairs. Future studies need to be done, to examine the consistency of these two approaches—using focal-paper keywords versus cited-reference keywords—in measuring topic-based novelty.

Acknowledgments

This work was financially supported by the Universities and Colleges Humanities and Social Sciences Benchmarking Project (Grant no. 113L9A001), Ministry of Education in Taiwan, and National Science and Technology Council, R.O.C. (Taiwan) under the grant numbers 113-2410-H-032-031-. The authors are also grateful to anonymous reviewers for their comments and suggestions.

References

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science, 62*(10), 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Bornmann, L., Tekles, A., Zhang, H. H., & Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics, 13*(1), 100979. https://doi.org/10.1016/j.joi.2019.100979

Boyack, K., & Klavans, R. (2014). Atypical combinations are confounded by disciplinary effects. In *Proceedings of the 19th International Conference on Science and Technology Indicators*. Leiden, The Netherlands.

Carayol, N., Agenor, L., & Oscar, L. (2019). The right job and the job right: Novelty, impact and journal stratification in science. *Impact and Journal Stratification in Science* (March 5, 2019).

Fontana, M., Iori, M., Montobbio, F., & Sinatra, R. (2020). New and atypical combinations: An assessment of novelty and interdisciplinarity. *Research Policy, 49*(7), 104063. https://doi.org/10.1016/j.respol.2020.104063

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review, 80*(5), 875–908. https://doi.org/10.1177/0003122415601618

Huang, Z., Tang, D., Zhao, R., & others. (2024). A scientific paper recommendation method using the time decay heterogeneous graph. *Scientometrics, 129*, 1589–1613. https://doi.org/10.1007/s11192-024-04933-4

Kanwal, T., & Amjad, T. (2024). Research paper recommendation system based on multiple features from citation network. *Scientometrics, 129*, 5493–5531. https://doi.org/10.1007/s11192-024-05109-w

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal, 36*(10), 1435–1457. https://doi.org/10.1002/smj.2294

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology, 57*(2), 251–263.

Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy, 44*(4), 684–697. https://doi.org/10.1016/j.respol.2014.10.007

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2*(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Ruan, X., Ao, W., Lyu, D., Cheng, Y., & Li, J. (2023). Effect of the topic-combination novelty on the disruption and impact of scientific articles: Evidence from PubMed. *Journal of Information Science.* https://doi.org/10.1177/01655515231161133

Tahamtan, I., & Bornmann, L. (2018). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics, 12*(3), 906–930. https://doi.org/10.1016/j.joi.2018.08.001

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008). ArnetMiner: Extraction and mining of academic social networks. *Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008)*, 990–998. https://doi.org/10.1145/1401890.1402008

Thagard, P. (2012). Creative combination of representations: Scientific discovery and technological invention. In R. Proctor & E. J. Capaldi (Eds.), *Psychology of science* (pp. 389–405). Oxford: Oxford University Press.

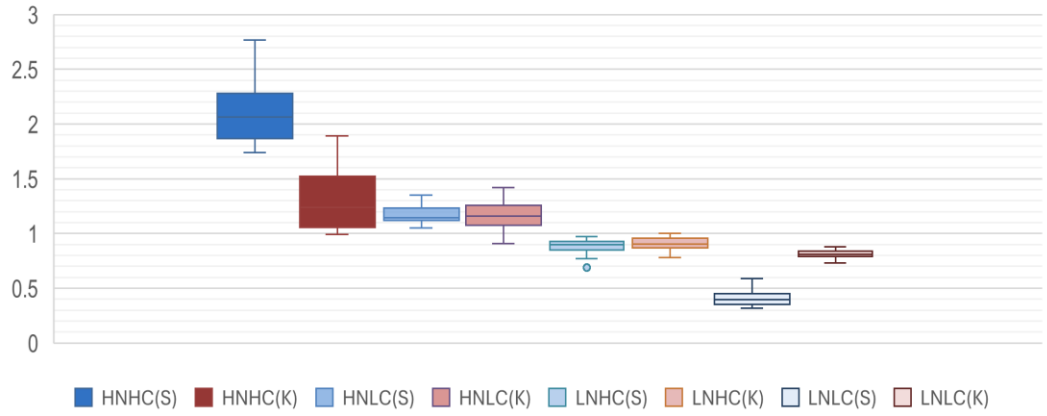
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*, 468-472. https://doi.org/10.1126/science.1240474

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013b). Supplementary materials for atypical combinations and scientific impact. *Science, 342*, 468. https://doi.org/10.1126/science.1240474

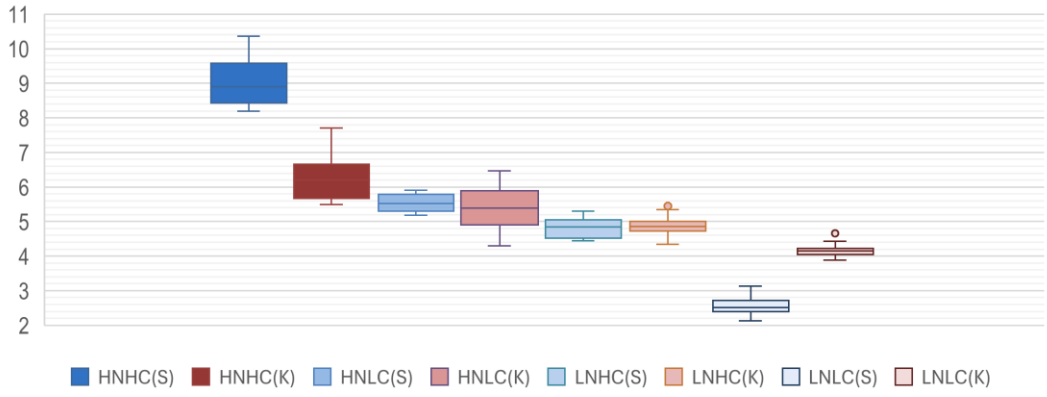
Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy, 46*(8), 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006

Zhang, F., & Wu, S. (2024). Predicting citation impact of academic papers across research areas using multiple models and early citations. *Scientometrics, 129*, 4137–4166. https://doi.org/10.1007/s11192-024-05086-0

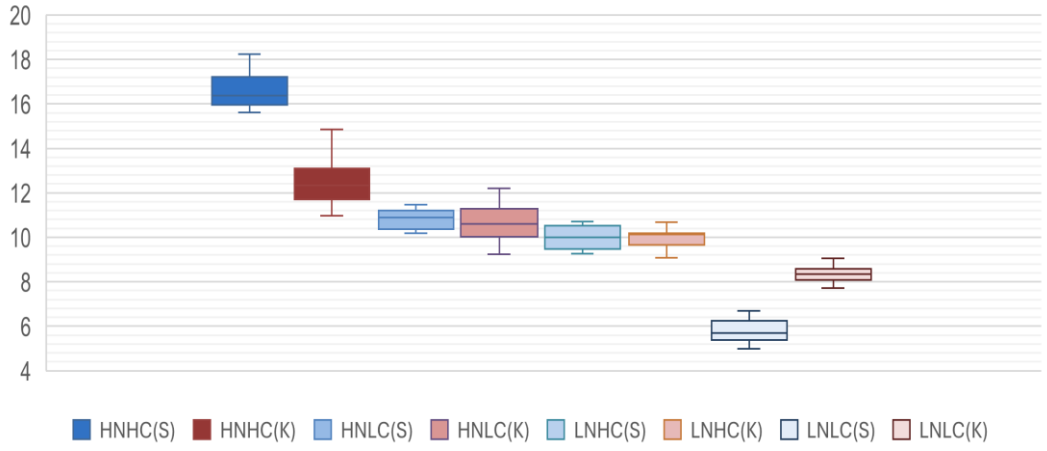
Appendix I



(a) Top 1% cited articles as highly-cited articles



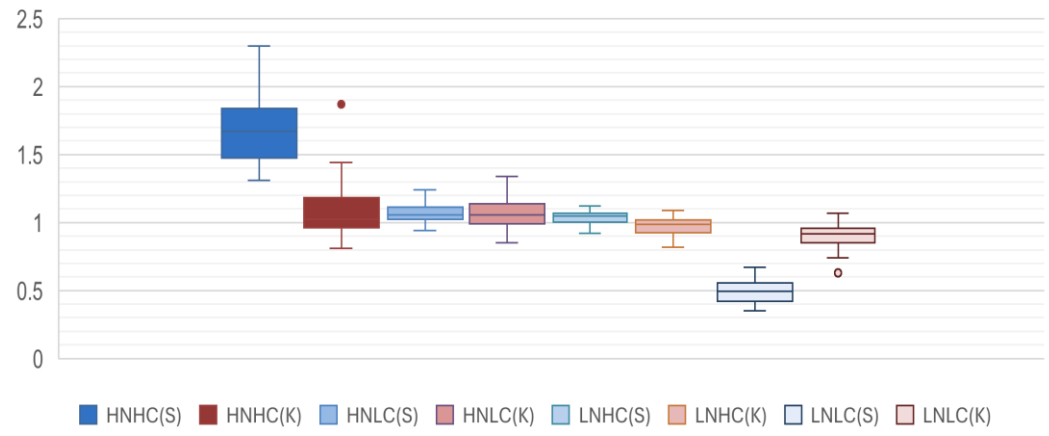
(b) Top 5% cited articles as highly-cited articles



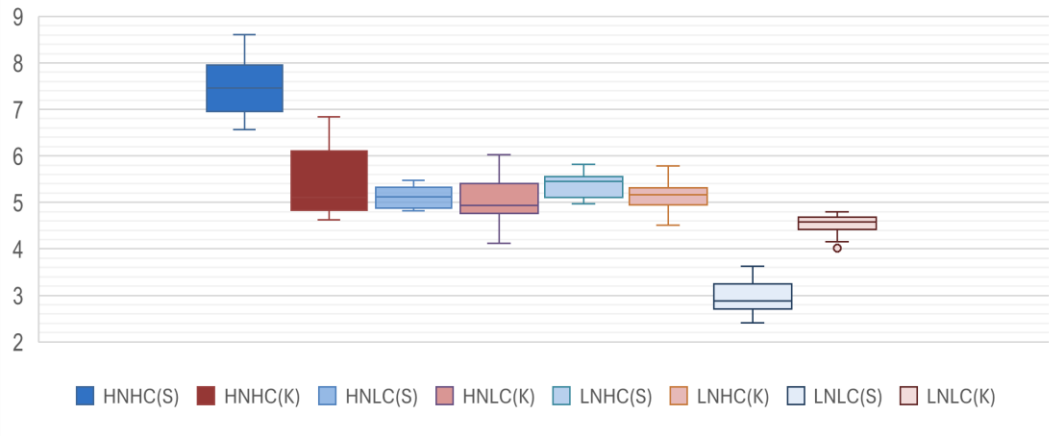
(c) Top 10% cited articles as highly-cited articles

Figure A1. Probabilities of being top highly cited across groups: Novelty (1). Y-axis represents the probability.

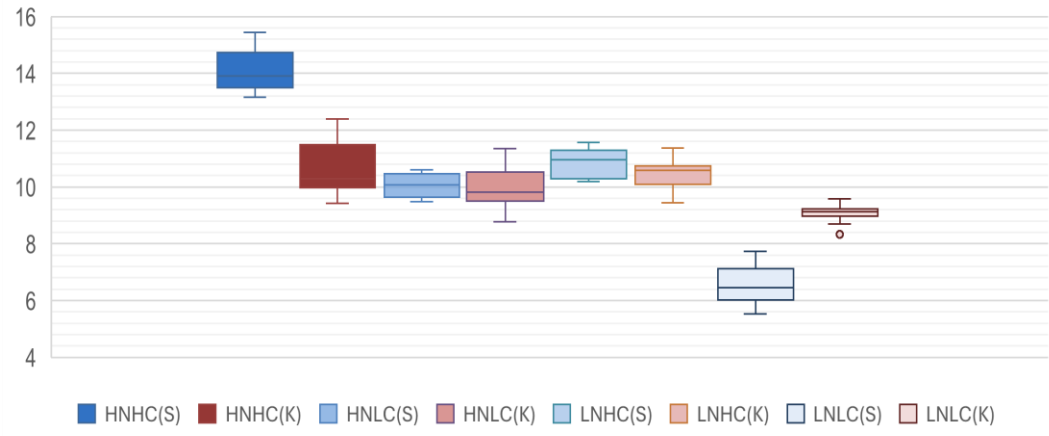
Appendix II



(a) Top 1% cited articles as highly-cited articles



(b) Top 5% cited articles as highly-cited articles



(c) Top 10% cited articles as highly-cited articles

Figure A2. Probabilities of being top highly cited across groups: Novelty (10). Y-axis represents the probability.

https://doi.org/10.51408/issi2025_078

Leveraging Large Language Models for Post-Publication Peer Review: Potential and Limitations

Mengjia Wu[1], Yi Zhang[2], Robin Haunschild[3], Lutz Bornmann[4]

[1]*Mengjia.Wu@student.uts.edu.au*, [2]*Yi.Zhang@uts.edu.au*
Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney (Australia)

[3]*R.Haunschild@fkf.mpg.de*,
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart (Germany)

[4]*L.Bornmann@fkf.mpg.de, bornmann@gv.mpg.de*
Max Planck Institute for Solid State Research, Heisenbergstr. 1, 70569 Stuttgart (Germany)
Science Policy and Strategy Department, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich (Germany)

Abstract

Peer review is the cornerstone of scientific evaluation, ensuring the quality, accuracy, and integrity of published research. However, challenges such as reviewer bias, time constraints, and the increasing volume of submissions have strained traditional peer review systems, resulting in delays, lower-quality reviews, and reviewer fatigue. These limitations highlight the need for innovative solutions. Large language models (LLMs) have emerged as promising tools to support or potentially replace certain aspects of peer review. This study investigates the potential of LLMs to enhance post-publication peer review, offering quality assessments and recommendations for published articles. Specifically, we designed two tasks to evaluate the performance of LLMs in post-publication research evaluation: identifying high-quality articles (Task 1) and providing ratings on recommended articles (Task 2). Six versions of three generative LLMs, including open-source models such as Qwen and Llama, the closed-source GPT-4o-mini model, and four BERT-based models, were assessed using in-context learning and fine-tuning approaches. The data for training and evaluation were sourced from H1 Connect (formerly Faculty Opinions), a platform for expert recommendations in the biomedical domain. Results indicate that fine-tuning LLMs with labelled data can significantly enhance their alignment with human expert evaluations. For Task 1, fine-tuned models performed well in identifying high-quality articles with an accuracy of 84%. However, for Task 2 - rating on recommended articles - LLMs struggled to match human judgement consistently with an accuracy below 0.6, highlighting their current limitations in nuanced, context-dependent tasks.

Introduction

In the realm of academic publishing, peer review serves as the cornerstone of scientific evaluation and dissemination (Bornmann, 2008). The process ensures that manuscripts meet certain standards of quality, accuracy, and integrity (defined by a certain field, community, journal etc.). Peer review, while essential, is not without challenges. Issues such as time constraints, reviewer biases (Bornmann, 2011), and the increasing volume of submissions necessitate solutions to enhance the efficiency and effectiveness of peer review. In this context, large language models (LLMs) have emerged as a promising tool for augmenting or replacing peer review. LLMs, exemplified by OpenAI's GPT series and Google's BERT, have

demonstrated remarkable capabilities in natural language understanding and generation (ChatGPT was introduced to the public in 2022, see Farhat et al., 2023). LLMs leverage vast amounts of textual data to learn linguistic patterns and generate human-like text. Their applications span various domains, including automated content generation, research classification (Wu et al., 2024), scholarly recommendation (Jia et al., 2025), knowledge association prediction (Wu et al., 2021), sentiment analysis, and language translation. More recently, the potential of LLMs to assist in research evaluation tasks has garnered attention from researchers and practitioners alike (Thelwall, 2024a, 2024b). LLMs have been used to undertake evidence synthesis and systematic assessment tasks (Joe et al., 2024), to propose references for anonymized in-text citations (Algaba et al., 2024), to predict citation counts, Mendeley reader counts, and social media engagement (de Winter, 2024; Vital Jr et al., 2024), and to identify prominent scholars (Sandnes, 2024).

The academic publishing landscape is witnessing significant growth (Bornmann et al., 2021), with an increasing number of manuscripts submitted for review and publication. The increasing number, while reflecting the importance of scientific inquiry for society, also places immense pressure on the peer review system. Reviewers and editors, as rule volunteers, face the task of evaluating numerous manuscripts and grant proposals within limited timeframes. Furthermore, the traditional peer review process has been often criticized for its subjectivity, potential biases, and the increasing difficulty in obtaining high-quality reviews. Consequently, delays in the review process, difficulties in finding reviewers, useless reports, and reviewer fatigue have become prevalent issues. These challenges highlight the need for innovative approaches to relieve the participants (reviewers) in the peer review process.

Several studies have explored the feasibility and effectiveness of using LLMs in peer review processes (Liang et al., 2024; Liu & Shah, 2023; López-Pineda et al., 2025; Thelwall & Yaghi, 2024). These studies suggest that LLMs can assist in specific peer review tasks such as identifying errors, verifying checklists, and providing feedback, but they are not yet reliable for complete evaluations of papers or proposals. One of these studies focused on the use of LLMs, specifically GPT-4, for specific reviewing tasks such as identifying errors, verifying checklists, and choosing the better paper among pairs of abstracts (Liu & Shah, 2023). The findings suggest that while LLMs can effectively identify errors and verify checklist questions with high accuracy, they struggle with more subjective tasks like discerning the quality of papers. This indicates that LLMs can serve as valuable assistants for specific, well-defined reviewing tasks but are not yet ready to replace human reviewers entirely.

Another empirical analysis evaluated the quality of feedback generated by GPT-4 on papers (Liang et al., 2024). The study compared LLM-generated feedback with human peer reviewer feedback across thousands of papers from prestigious journals and conferences. The results show a significant overlap between the points raised by GPT-4 and human reviewers, particularly for weaker papers. An additional user study revealed that researchers found the LLM-generated feedback helpful, suggesting that LLMs can provide valuable assistance in the peer review process,

especially for researchers in under-resourced settings. The most recent study (Thelwall & Yaghi, 2024) evaluated whether ChatGPT 4o-mini can estimate the quality of papers by comparing its scores to departmental averages across 34 Units of Assessment in the United Kingdom's Research Excellence Framework (REF) 2021. The results show a generally positive correlation, with some variations, suggesting that LLMs can provide reasonable quality estimates, especially in the physical and health sciences. These assessments are based only on titles and abstracts, not comprehensive evaluations.

The previous studies on the use of LLMs in the peer review process reveal that their use holds significant promise for addressing some of the challenges associated with traditional peer review. Although LLMs may provide valuable feedback, it is essential to recognize their limitations. For example, LLMs seem to include "hallucinating" information into otherwise plausible responses (Thelwall, 2024b). Ongoing research should try to refine these models to ensure their effective and ethical use in the academic community. Building on the insights from previous studies, the current empirical investigation aims to evaluate the use of LLMs for post-publication peer review. Post-publication review, unlike traditional pre-publication review, occurs after the paper has been published, providing a platform with recommendations and quality assessments of papers. This study seeks to assess the opportunities of LLMs in enhancing post-publication peer review processes. By leveraging advanced LLMs, the study aims to explore how these models may complement human expertise and streamline the review workflow.

In this study, we designed two tasks to assess the LLMs' capabilities in post-publication research evaluation: identifying high-quality articles (Task 1) and recommended article rating (Task 2). Six versions of generative LLMs, including open-source Qwen, Llama models, and closed-source GPT-4o-mini model, in addition with four BERT-based language models, were tested under two different learning settings: in-context learning and fine-tuning, to complete the two tasks. Using data from H1 Connect (a post-publication peer review service in medicine and life sciences, formerly known as Faculty Opinions) as training and test data, we performed model comparisons on both tasks. The results revealed that, with an appropriate fine-tuning strategy, current LLMs have strong potential to serve as preliminary reviewers to identify high-quality papers (Task 1), with the fine-tuned GPT-4o-mini model achieving the accuracy of 84% and BERT models above 75%. However, the models still lack the capabilities to achieve expert-level judgment when facing more complicated tasks like article rating (Task 2), in which rating differences are more nuanced to learn.

Data and Tasks

Data source

H1 Connect is a specialized platform designed to provide expert recommendations and support research evaluation in the biomedical domain. It delivers scholarly output metadata along with expert-generated recommendations, which are enriched with detailed ratings, commentaries, and classification codes. The additional

information explains the basis for the inclusion of the papers on the platform and their relevance for the community. We selected the H1 Connect data for its extensive data coverage and rich evaluation metadata across biomedical fields, which ensures a representative and diverse dataset for comparing assessments from experts and other instruments such as bibliometrics or LLMs.

Task formulation

To examine the research evaluating capabilities of LLMs, we designed two tasks of high-quality article identification (Task 1) and recommended article rating (Task 2). To achieve the tasks, we collected two datasets from H1 Connect, with their details given in descriptions below and Table 1. Given that testing LLMs on the global dataset comes with an unneglectable burden of computational costs, we randomly sampled partial articles for each task from the entire dataset. We used the article abstracts as our input to the models due to the incomplete availability of full texts.

Task 1 - High-quality article identification: This task aims to evaluate how effectively LLMs can identify high-quality articles from a mixed pool of high- and low-quality articles, compared to the judgment of human experts. Low-quality articles are defined as those with no expert recommendations, and high-quality articles are those with three or more expert recommendations. To construct a mixed pool for testing, we compiled 4,538 articles from OpenAlex (Priem et al., 2022) – a bibliographic catalogue of scientific papers – with no expert recommendations and 4,994 articles with three or more expert recommendations. The not-recommended articles were published between 2010 and 2020 in the same journal, with the same volume and issue as the recommended papers. We excluded the journals *Science*, *Nature*, *Proceedings of the National Academy of Sciences of the United States of America*, *Science Advances*, *Nature Communications*, *Scientific Reports*, and *PLOS ONE* due to their multidisciplinary nature for the selection of not-recommended papers. The selected LLMs are required to retrieve the 4,538 high-quality articles from this pool as accurately as possible.

Task 2 - Recommended article rating: This task delves into a more detailed objective of rating research articles based on their quality and content. To avoid complications in synthesizing expert ratings, we focused on articles with only one recommendation at this stage. The data collection also follows procedures as in Task 1, resulting in 86,805 articles with a rating of 1, 54,154 articles with a rating of 2, and 11,089 articles with a rating of 3 (roughly 8:5:1). Considering the computational costs for model testing, we sampled a balanced dataset that consists of 5,000 articles from each rating of 1 (good), 2 (very good), and 3 (excellent), ending with 15,000 articles.

Table 1. Descriptions of datasets used in Task 1 and Task 2.

Task 1	# Article	**Three recommendations**	**No recommendation**	
		4,994	4,538	
Task 2	# Article	**1 (Good)**	**2 (Very good)**	**3 (Exceptional)**
		5,000	5,000	5,000

Methodology framework

The overall research framework is presented in Figure 1. To perform Task 1 and Task 2, we selected four BERT variant models and six generative LLMs, with details provided in the model selection section. Two representative model adaptation techniques, in-context learning (ICL) and fine-tuning, were employed to adapt the models to output the desired results for the tasks. These techniques are described in detail in the following subsections.

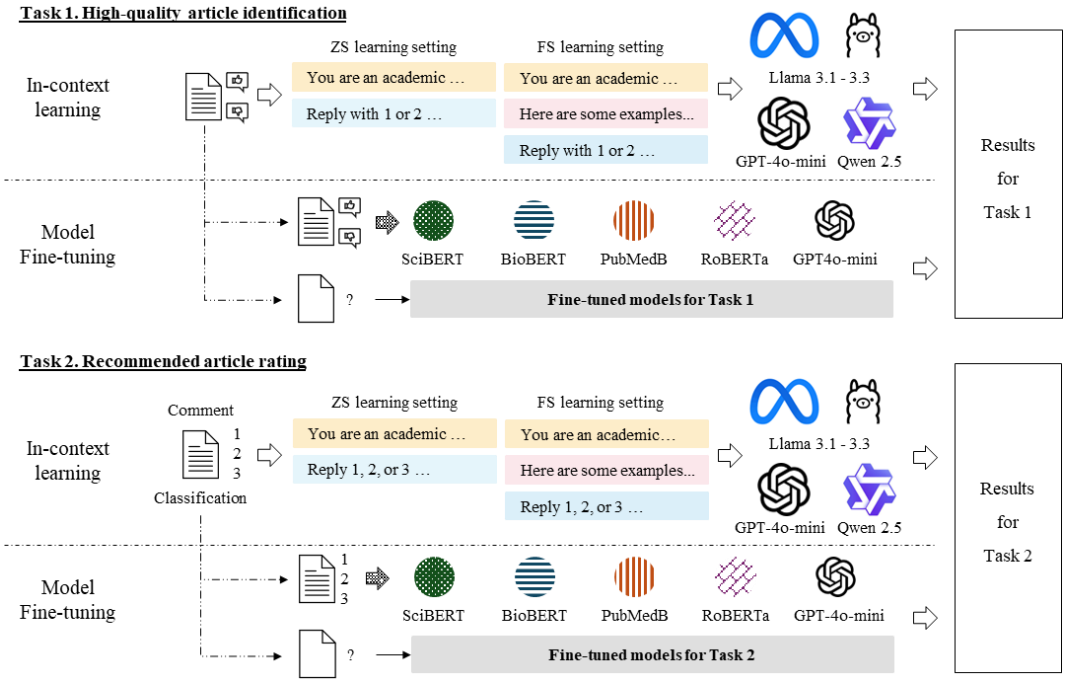


Figure 1. The overall research framework.

Model selection

Four BERT variant models: SciBERT (Beltagy et al., 2019), BioBERT (Lee et al., 2020), RoBERTa (Liu et al., 2019), and PubMedBERT (Gu et al., 2021) are encoder-only language models built on the transformer architecture, which converts the input language as embeddings for downstream analysis. The key distinction among these models lies in their training corpora and methods. SciBERT is tailored for scientific NLP tasks, pre-trained on 1.14 million scientific articles from

Semantic Scholar[1]. BioBERT extends the original BERT pretraining corpus by incorporating 29 million PubMed abstracts and full-text articles from PubMed Central[2], enhancing its performance in the biomedical domain. PubMedBERT also targets biomedical domain, but it exclusively uses PubMed abstracts and PubMed Central full-text articles for pretraining, omitting the general BERT corpus, which makes it more specialized for biomedical tasks. RoBERTa, a refined version of BERT, optimizes the pretraining procedures with modified training parameters and task settings, improving model efficiency and performance while retaining general-purpose applicability.

Current generative LLMs generally employ decoder-only architecture, enabling them to generate text sequences directly based on the given natural language input. The widespread adoption of ChatGPT has shown the remarkable capabilities of such models in language comprehension, text generation, and question-answering. Apart from GPT models, multiple big tech companies have developed and released open-source models for public access and use, represented by Llama models from Meta (formerly Facebook) and Qwen models from Alibaba. Given that, we selected multiple representative open- and closed-source models considering computing budget and time costs. For open-source models, we intentionally chose both the smallest (3B or 7B, in which B indicates billion parameters) and largest versions (70B or 72B) to test how the model size can affect evaluation results. The tested models in the final pool include: GPT-4o-mini (Achiam et al., 2023) from OpenAI, Llama 3.1-8B, Llama 3.2-3B, and Llama 3.3-70B from Meta (Dubey et al., 2024), as well as Qwen 2.5-7B and Qwen 2.5-72B from Alibaba (Yang et al., 2024).

ICL for generative LLMs

ICL is a prompt-engineering technique designed for generative LLMs (GPT-4o-mini, Llama, and Qwen models in this paper). ICL works by providing contextual information, sometimes along with task-specific input-output pair demonstrations directly in the prompts, enabling models to generate responses for given questions. Unlike fine-tuning, ICL does not alter the model's parameters; instead, it modifies the prompts to achieve more accurate outputs. This makes ICL a low-cost and user-friendly approach to leveraging LLMs. In this study, we employed two of the most prevalent ICL prompting schemes:

- Zero-shot (ZS) learning setting: In the ZS setting, the prompt only includes descriptions of the task as contextual information. The LLMs generate recommendations (Task 1) or ratings (Task 2) for each article without any additional contextual information.
- Few-shot (FS) learning setting: In the FS setting, the prompt includes both the task description and five demonstrations of input-output pairs (see the Supplementary Material) for each class. For Task 1, five recommended articles and five non-recommended articles, along with their abstracts and expert recommendations, are provided. For Task 2, five articles from each rating

[1] https://www.semanticscholar.org
[2] https://pubmed.ncbi.nlm.nih.gov

category (1, 2, and 3) are presented with their ratings. The demonstrations are selected randomly, and each inference is conducted using a different set of demonstrations.

ICL is an idealized learning setting that anticipates LLMs to complete the tasks accurately with the given contextual information (task description) or a few samples. We designed three sets of prompt templates (p1-p3) for Task 1 and Task 2 to instruct generative LLMs. The prompts and their corresponding usage for each task are provided in the Supplementary Material.

Language model fine-tuning

Fine-tuning is a model retraining method that adapts LLMs to specific tasks by updating their parameters using labelled data (in our case, the labels are article recommendations and ratings). Unlike training from scratch, fine-tuning can retain knowledge learnt during the pre-training stage in the retraining process. However, compared to ICL, fine-tuning, especially for generative LLMs, is much more computationally intensive. Additionally, fine-tuned models tend to be more task-specific, which may reduce their generalizability. This strategy can be applied to both BERT models and generative LLMs. Due to the high computational costs of fine-tuning the selected generative LLMs on local machines, we only applied this learning setting for BERT models and the GPT-4o-mini model (through the OpenAI API).

Validation metrics

Four validation metrics were employed to measure the models' performance in Task 1 and Task 2. The definitions and calculations are given as follows:

- Accuracy (A): Accuracy measures the ratio of correctly classified articles to all articles.
- Precision (P): For a specific category, P is the ratio of correctly classified articles to all articles predicted as positive for that class.
- Recall (R): For a specific category, R is the ratio of correctly classified articles to all articles that belong to that class.
- Cohen's kappa coefficient ($\boldsymbol{\kappa}$): κ measures the level of agreement between a LLM and a human expert on the classification task. It ranges from -1 to 1, with the larger value indicating higher agreement.

$$\kappa = \frac{A - p_e}{1 - p_e}$$

$$p_e = \frac{1}{N}\sum_k n_L^k n_H^k$$

N is the total number of articles and k is the number of categories to be classified (recommended or not recommended in Task 1, rating of 1, 2, or 3 in Task 2), n_L^k and n_H^k, respectively, denote the number of articles classified to category k by LLMs (L) and human experts (H). Landis and Koch (1977) characterize values < 0

as indicating no agreement and 0-0.20 as slight agreement, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement.

Results

Results for high-quality article identification (Task 1)

For the ICL strategies, we tested all generative LLMs on all articles in Task 1 (a total of 9,532 articles). The predictive results are shown in Figures 2 and 3, where the green and red areas represent the outputs recommended and not recommended respectively, and deep and light colors refer to articles correctly or wrongly classified (the sum of each bar may be slightly smaller than 9,532 due to a few invalid answers from LLMs).

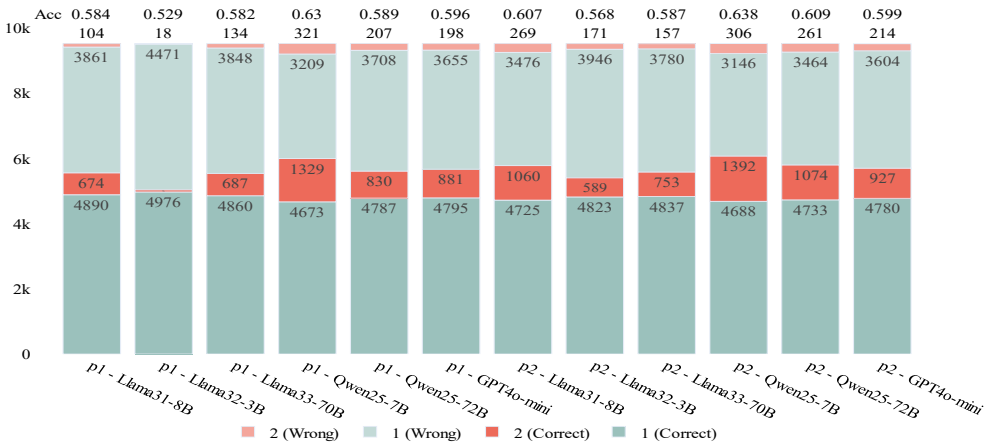


Figure 2. Model results using the ZS learning setting in Task 1.

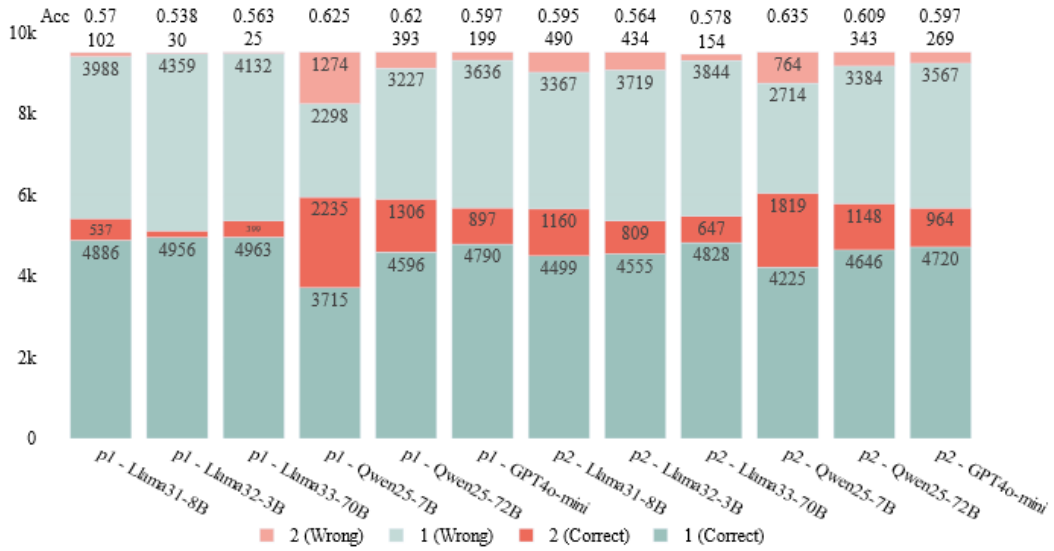


Figure 3. Model results using the FS learning setting in Task 1.

In both Figures 1 and 2, p1 refers to the prompt without evaluation criteria details given, and p2 refers to the prompt with evaluation criteria details (see the Supplementary Material). The accuracy, precision, and recall metrics for the generated answers are provided in Table 2.

Under ICL settings, the overall accuracy of tested LLMs is around 0.6, which is barely satisfying for a binary classification task. It can be observed from Figures 2 and 3 that most LLMs are inclined to generate biased positive answers (recommended) for articles – even though half of them did not receive any recommendations from human experts. This tendency is also reflected in the generally low recall rate for the "not recommended" class in Table 2. Besides, the performance of the closed-source model, GPT-4o-mini, does not show significant advancements compared to other open-source language models.

Despite that, the model outputs are also subject to which prompt and what learning setting were used. In Figures 2 and 3, the accuracies of most models increase when changing the prompt from p1 to p2, i.e., using more detailed evaluation criteria in the prompt. Details of the evaluation criteria are essentially critical for LLMs to give more accurate justification for article recommendations.

However, switching from ZS to FS setting, i.e., providing some examples to LLMs, does not let LLMs make more accurate recommendations. It increased the ratio of articles predicted as "not recommended", but the accuracy did not improve accordingly. In other words, showing both positive and negative samples, i.e., articles recommended and not recommended by human experts to LLMs, can help them to produce more critical opinions, but the alignment with human experts still struggles. This indicates that article evaluation can be a complex and long content-dependent task – realizing human-level judgment may still require a deeper understanding of articles than a few examples can provide.

When comparing results from smaller versions of models to larger versions under ICL settings, the accuracies did not show significant improvements – in most cases, the accuracy dropped slightly. Although it has been proven that larger models can perform significantly better in most generalized tasks (Touvron et al., 2023; Yang et al., 2024), our results indicate that model size is not a decisive factor in this pure binary classification task of differentiating recommended and not recommended articles under ICL settings.

The results of the fine-tuned models are presented in Table 2. Under the fine-tuning learning setting, both generative LLMs and BERT models are retrained to learn patterns for recommending articles from labelled data and then used to predict unseen records. We split the dataset into an 80% training set and a 20% test set. The optimal learning rate and number of training epochs were empirically determined by monitoring the training and validation loss.

Table 2. LLM results for Task 1 under ZS, FS, and fine-tuning learning settings*

Setting	Prompt	Model	A	κ	P (Y)	R (Y)	P (N)	R (N)
ICL	ZS	p1	Llama 3.1-8B	0.584	0.133	0.559	0.979	**0.866**	0.149
			Llama 3.2-3B	0.529	0.012	0.527	**0.996**	0.788	0.015
			Llama 3.3-70B	0.582	0.13	0.558	0.973	0.837	0.151
			Qwen 2.5-7B	0.630	0.235	0.593	0.936	0.805	0.293
			Qwen 2.5-72B	0.589	0.147	0.564	0.959	0.8	0.183
			GPT-4o-mini	0.596	0.160	0.567	0.960	0.816	0.194
		p2	Llama 3.1-8B	0.607	0.186	0.576	0.946	0.798	0.234
			Llama 3.2-3B	0.568	0.099	0.55	0.966	0.775	0.130
			Llama 3.3-70B	0.587	0.14	0.561	0.969	0.827	0.166
			Qwen 2.5-7B	**0.638**	**0.253**	**0.598**	0.939	0.82	**0.307**
			Qwen 2.5-72B	0.609	0.191	0.577	0.948	0.804	0.237
			GPT-4o-mini	0.599	0.168	0.57	0.957	0.812	0.205
	FS	p1	Llama 3.1-8B	0.57	0.102	0.551	0.980	0.84	0.119
			Llama 3.2-3B	0.538	0.031	0.532	0.994	0.845	0.036
			Llama 3.3-70B	0.563	0.087	0.546	**0.995**	**0.941**	0.088
			Qwen 2.5-7B	0.625	0.24	**0.618**	0.745	0.637	**0.493**
			Qwen 2.5-72B	0.620	0.215	0.587	0.921	0.769	0.288
			GPT-4o-mini	0.597	0.164	0.568	0.96	0.818	0.198
		p2	Llama 3.1-8B	0.595	0.163	0.572	0.902	0.703	0.256
			Llama 3.2-3B	0.564	0.095	0.551	0.913	0.651	0.179
			Llama 3.3-70B	0.578	0.118	0.557	0.969	0.808	0.144
			Qwen 2.5-7B	**0.635**	**0.253**	0.609	0.847	0.704	0.401
			Qwen 2.5-72B	0.609	0.19	0.579	0.931	0.77	0.253
			GPT4o-mini	0.597	0.164	0.57	0.946	0.782	0.213
Fine-tuned on the training set (80% data)			SciBERT	0.785	0.564	0.764	0.863	0.817	0.696
			BioBERT	0.789	0.574	0.778	0.845	0.804	0.725
			RoBERTa	0.761	0.512	0.726	**0.885**	0.825	0.62
			PubMedBERT	0.802	0.599	0.784	0.866	**0.827**	0.728
			GPT-4o-mini	**0.84**	**0.679**	**0.878**	0.811	0.802	**0.872**

* Note: Results in bold font indicate the best accuracy, underlined results are the second best. We separated the comparison by experimental settings (ZS, FS and fine-tuning).

The fine-tuned GPT-4o-mini achieved the highest accuracy among all models, including the fine-tuned BERT models, which utilize encoder-only architectures optimized for tasks like text understanding and classification rather than generation. This result highlights the superiority of larger-scale LLMs in handling versatile

tasks and supports the scaling law in language models (Kaplan et al., 2020), which suggests that model performance improves to some extent with increasing size. BERT models typically have around 110 million parameters, while GPT models often utilize models with billions of parameters.

To compare the inter-model agreement on Task 1, we depicted the heatmap based on the pairwise κ of model outputs in Figure 4 – the darker the red, the higher the agreement is between the models. The overall agreement with human experts is the same as reflected by accuracy: Fine-tuned models are generally above 0.6 but models under the ICL settings are all lower than 0.3. Regarding the inter-model agreement, fine-tuned models show satisfying moderate agreements above 0.6, following the interpretation of Landis and Koch (1977). Notably, some models under the ICL settings also exhibit good inter-model agreement (above 0.6), including Qwen 2.5-72B, GPT-4o-mini, and Llama 3.3-70B, which are all LLMs in their larger versions. These results indicate that larger models may have more consistent behaviors when dealing with the less complicated Task 1. The results should be interpreted against the backdrop of results on the agreement of reviewers from the (pre-publication) peer review process. The results of a meta-analysis of Bornmann et al. (2010) including several primary journal peer review studies show that the agreement between reviewers assessing the same manuscript is low (in general): The pooled κ across 48 studies is 0.17. The results for the agreement of human experts and models are relatively high in this study compared to the results from the meta-analysis.

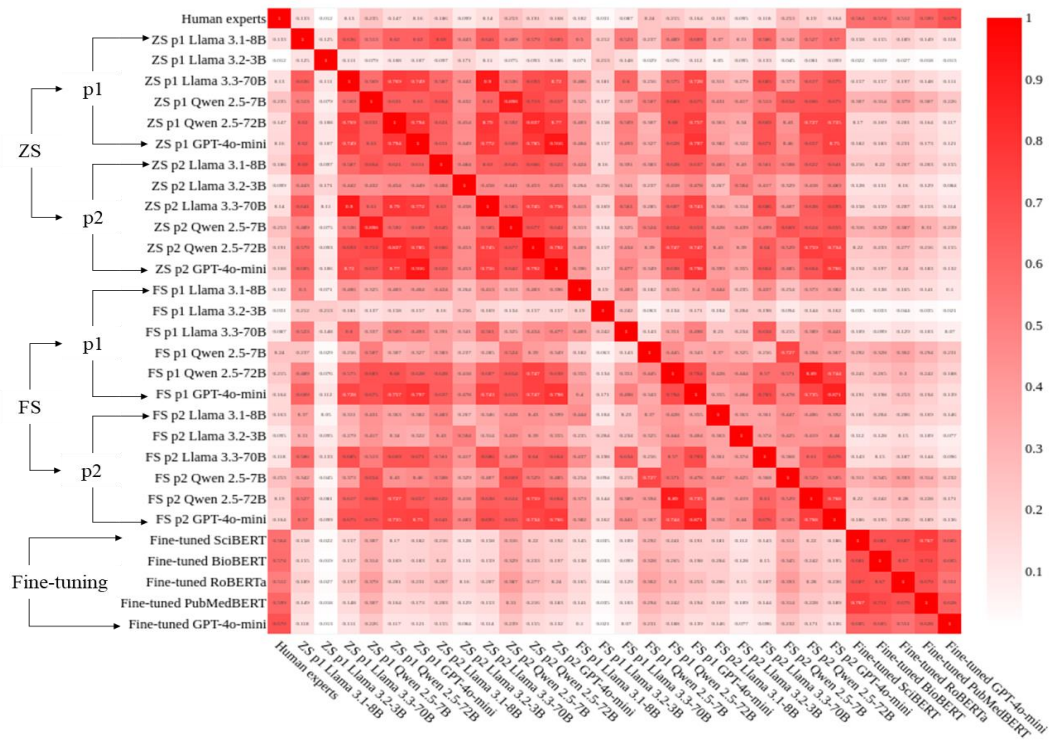


Figure 4. The heatmap of κ between LLM outputs in Task 1.

Results for recommended article rating (Task 2)

In Task 2, prompt p3 (see the Supplementary Material) was used to instruct LLMs to give ratings of 1, 2, and 3 for each article provided. The results are presented in Figure 5 and Table 3. In Figure 5, the colors represent three different ratings: Green – 1, Red – 2, and Yellow – 3 (the sum of each bar may be slightly smaller than 15,000 due to a few invalid answers from LLMs). The overall low accuracy below 0.4 highlights the challenge of differentiating article ratings under the ICL settings. Among the models, Qwen 2.5-72B achieved the highest accuracy but still presented a relatively biased preference for ratings of 2 and 3. Llama 3.1-8B within the FS setting yielded rather balanced predictions but suffered from lower accuracy. The other models, excluding Llama 3.1-8B and Llama 3.2-3B models under the FS setting, tend to show the inclination to ratings of 2 and 3.

Unlike Task 1, switching from the ZS to the FS setting significantly altered the outputs of most models, but the direction of this change depends on which specific model is used: Llama 3.1-8B produced much more balanced results with the few samples provided, Llama 3.2-3B changed its main preference from ratings of 2 to 1, results from Llama 3.3-70B did not change much, FS increases the number of ratings of 2 and 3 for Qwen 2.5-72B and GPT-4o-mini. However, the accuracy of all model outputs still did not improve much. Despite those changes, the results endorse our previous claim in Task 1: The regular FS learning setting is not an effective learning strategy for research evaluation tasks.

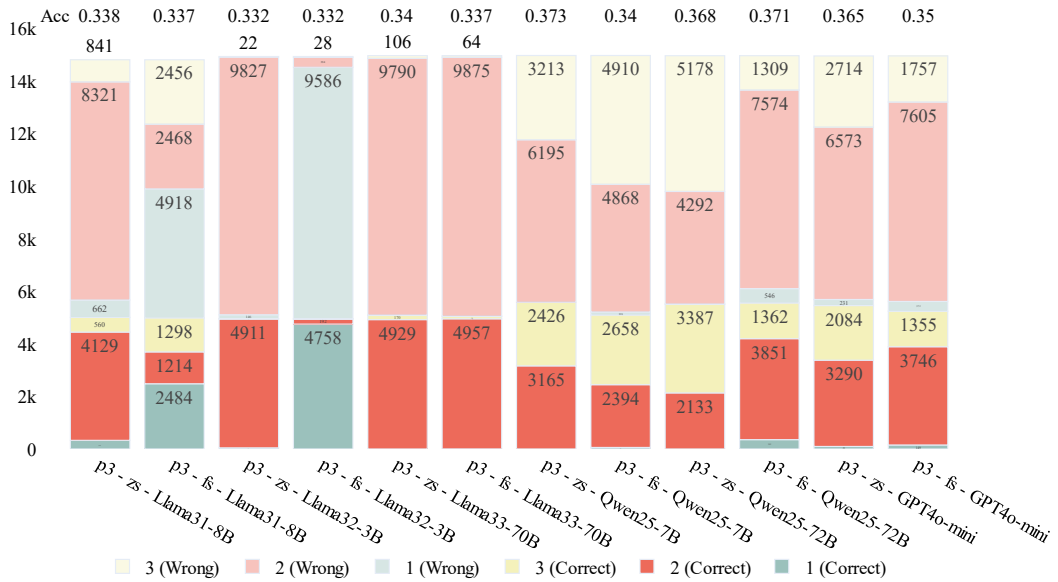


Figure 5. LLM results for Task 2 under ZS and FS learning settings.

Table 3. LLM results for Task 2 under ZS, FS, and fine-tuning settings*

	Setting	Model	A	κ	P1*	R1*	P2	R2	P3	R3
ICL	ZS	Llama 3.1-8B	0.334	0.007	0.329	**0.065**	0.332	0.826	0.4	0.112
		Llama 3.2-3B	0.332	0.001	0.22	0.008	0.334	0.985	0.421	0.003
		Llama 3.3-70B	0.34	0.01	<1e-3	<1e-3	0.335	**0.986**	**0.616**	0.034
		Qwen 2.5-7B	**0.373**	**0.059**	**1**	<1e-3	**0.338**	0.633	0.43	0.485
		Qwen 2.5-72B	0.368	0.052	0.2	<1e-3	0.332	0.427	0.396	**0.678**
		GPT-4o-mini	0.364	0.047	0.286	0.019	0.333	0.658	0.434	0.417
	FS	Llama 3.1-8B	0.336	0.005	0.335	0.502	0.329	0.245	0.347	0.262
		Llama 3.2-3B	0.331	0.002	0.332	**0.952**	0.333	0.038	0.349	0.003
		Llama 3.3-70B	0.337	0.006	**0.750**	0.001	0.334	**0.991**	**0.602**	0.019
		Qwen 2.5-7B	0.34	0.011	0.318	0.011	0.33	0.479	0.351	**0.532**
		Qwen 2.5-72B	**0.371**	**0.057**	0.392	0.07	**0.337**	0.77	0.51	0.272
		GPT-4o-mini	0.35	0.025	0.282	0.03	0.33	0.749	0.436	0.271
Fine-tuning	Test set	SciBERT	0.453	0.176	0.466	0.621	0.344	0.263	0.525	0.463
		BioBERT	0.458	0.182	0.459	0.68	**0.361**	0.208	0.515	0.47
		RoBERTa	0.452	0.172	0.442	**0.719**	0.357	0.162	0.518	0.455
		PubMedBERT	0.461	0.187	0.464	0.68	**0.361**	0.231	0.527	0.456
		GPT-4o-mini	**0.463**	**0.195**	**0.533**	0.466	0.348	0.395	0.527	0.526
	Extra test set	SciBERT	0.493	0.111	0.629	0.712	0.394	0.186	0.146	0.349
		BioBERT	0.499	0.122	0.627	0.716	0.418	0.183	0.162	**0.402**
		RoBERTa	0.492	0.098	0.616	**0.728**	0.383	0.156	0.153	0.357
		PubMedBERT	0.512	0.14	0.635	0.715	**0.453**	0.23	0.158	0.361
		GPT-4o-mini	**0.561**	**0.165**	**0.653**	0.682	0.431	**0.493**	**0.444**	0.036

* Note: P1 and R1 respectively refers to the precision and recall of category 1. Results in bold font indicate the best accuracy, underlined results are the second best. We separated the comparison by experimental settings (ZS, FS and fine-tuning).

In addition to the standard test set, we created an extra test set for Task 2 to validate the performance of the fine-tuned models in real-world settings. The new test set is a dataset that simulates the imbalanced distribution of ratings in real-world scenarios – containing 1,675 records of rating 1, 1,076 records of rating 2, and 249 records of rating 3. This corresponds roughly to a 8:5:1 ratio as introduced in the full dataset we collected.

The results indicate that the fine-tuned GPT-4o-mini achieved the overall best performance on both test sets, especially on the extra real-world simulated test set. The second best-performing fine-tuned model is PubMedBERT, the BERT variant

trained specifically on PubMed articles corpora. Generally, all the language models tested on this task presented an accuracy lower than 0.6 and a κ agreement with human experts below 0.2. The low measures indicate that in Task 2, the differences between the three ratings are much more nuanced than in Task 1. It seems that Task 2 is more challenging for language models to learn different articles' quality based on their abstracts.

The inter-model κ agreement of Task 2 is visualized in Figure 6. Compared to Task 1, the agreement among fine-tuned models and models under ICL settings both dropped to lower than 0.6 and 0.2. Despite generally low agreement of LLMs under ICL settings, Qwen 2.5-72B and GPT-4o-mini still showed relatively high agreement with each other.

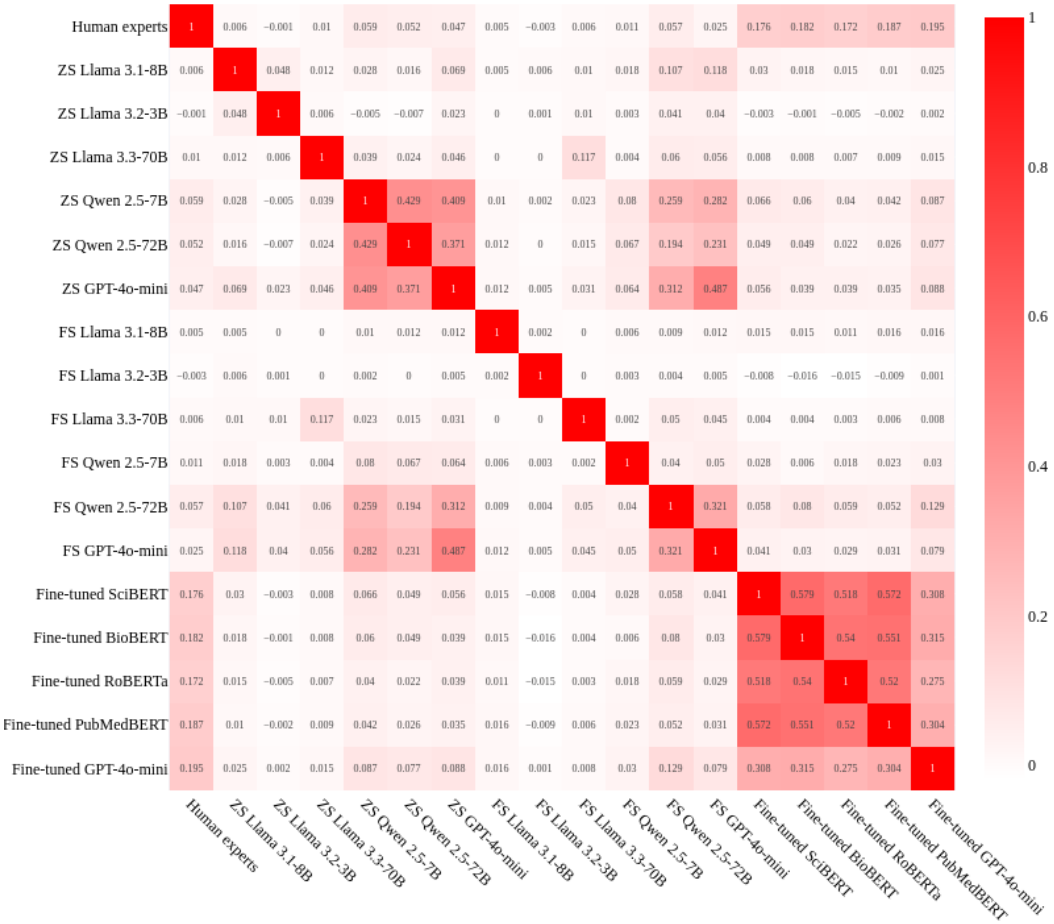


Figure 6. The heatmap of κ between LLM outputs in Task 2.

Conclusions

In this study, we performed a thorough comparison of current LLMs' performance on research evaluation tasks under ICL (ZS and FS) and fine-tuning learning settings, providing insights into leveraging LLMs for post-publication review and rating. Overall, our results demonstrate that LLMs fine-tuned with partial human expert annotations can serve as a preliminary tool for initial research evaluation. However, more complicated tasks, like rating articles on a specific scale, are more challenging and may require more resources and sophisticated methodologies. More specifically, the key findings of this study are as follows:

- Among the three model learning settings, fine-tuning works significantly better and aligns with expert opinions the most, but this comes with a trade-off of requiring a certain amount of existing training data. The idealized settings of utilizing LLMs, like ZS and FS, which anticipate LLMs to perform evaluation independently or with very limited contextual information, are still compromised in their alignment with human experts in real-world practice.
- Among the fine-tuned models, GPT-4o-mini is the best among the tested LLMs, including BERT-based models and open-sourced generative LLMs.
- Under the fine-tuning setting, LLMs can offer relatively satisfying performance on identifying high-quality articles (Task 1) with very little training data but may struggle to accurately rate recommended articles (Task 2). The selected LLMs, even after fine-tuning, are still prone to giving biased answers that are different from those of human experts.

Limitations and future directions

Certain limitations come with this work. First, we did not apply fine-tuning strategies on open-source LLMs like Qwen and Llama due to the restraints from high computational resource requirements, leading to the lack of comparison of those options in our study. Second, in this paper, we only fed article abstracts to LLMs for evaluation, which contain very concise and limited information and may be insufficient for evaluating the overall quality of research articles. Third, LLMs are the only knowledge sources for performing research evaluation tasks. No external data sources, which can be academic knowledge graphs containing more enriched information, have been leveraged. Aiming to equip LLMs with better capabilities and accuracy of research evaluation, the future directions of this study will spread to three perspectives: (1) employ more computational resources to realize fine-tuning on open-source LLMs, (2) develop a work pipeline for multi-modal LLMs to systematically process article full texts with figures and tables affiliated, and (3) incorporate external data resources with LLMs to realize enriched context-aware evaluation.

Acknowledgments

Mengjia Wu and Yi Zhang are supported by the Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, in conjunction with the National Science Foundation (NSF) of the United States, and CSIRO-NSF #2303037. We would like to thank H1 staff for providing data access. Access to OpenAlex bibliometric data has been supported via the German Competence Network for Bibliometrics (Schmidt et al., 2024), funded by the Federal Ministry of Education and Research (grant number: 16WIK2101A).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). *GPT-4 technical report* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.08774

Algaba, A., Mazijn, C., Holst, V., Tori, F., Wenmackers, S., & Ginis, V. (2024). *Large language models reflect human citation patterns with a heightened citation bias* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2405.15739

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A pretrained language model for scientific text* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1903.10676

Bornmann, L. (2008). Scientific peer review. An analysis of the peer review process from the perspective of sociology of science theories. *Human Architecture - Journal of the Sociology of Self-Knowledge*, *6*(2), 23-38. http://www.okcir.com/HAVI2SPRING2008.html

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*, 199-245. https://doi.org/10.1002/aris.2011.1440450112

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 224. https://doi.org/10.1057/s41599-021-00903-w

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PloS one*, *5*(12), e14331.

de Winter, J. (2024). Can ChatGPT be used to predict citation counts, readership, and social media interaction? An exploration among 2222 scientific abstracts. *Scientometrics*, *129*(4), 2469-2487. https://doi.org/10.1007/s11192-024-04939-y

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., & Fan, A. (2024). *The llama 3 herd of models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.21783

Farhat, F., Sohail, S. S., & Madsen, D. Ø. (2023). *How trustworthy is chatGPT? The case of bibliometric analyses*. Retrieved August, 19 from https://doi.org/10.20944/preprints202303.0479.v1

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, *3*(1), 1-23.

Jia, R., Wu, M., Ding, Y., Lu, J., & Zhang, Y. (2025). *HetGCoT-Rec: Heterogeneous Graph-Enhanced Chain-of-Thought LLM Reasoning for Journal Recommendation* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.01203

Joe, E. T., Koneru, S. D., & Kirchhoff, C. J. (2024). *Assessing the effectiveness of GPT-4o in climate change evidence synthesis and systematic assessments: Preliminary insights*. Retrieved July, 23 from https://arxiv.org/abs/2407.12826

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2001.08361

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., & Yin, Y. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, *1*(8), AIoa2400196.

Liu, R., & Shah, N. B. (2023). *Reviewergpt? an exploratory study on using large language models for paper reviewing* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2001.08361

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1907.11692

López-Pineda, A., Nouni-García, R., Carbonell-Soliva, Á., Gil-Guillén, V. F., Carratalá-Munuera, C., & Borrás, F. (2025). Validation of large language models (Llama 3 and ChatGPT-4o mini) for title and abstract screening in biomedical systematic reviews. *Research Synthesis Methods*, 1-11.

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2205.01833

Sandnes, F. E. (2024). Can we identify prominent scholars using ChatGPT? *Scientometrics*, *129*(1), 713-718. https://doi.org/10.1007/s11192-023-04882-4

Schmidt, M., Rimmert, C., Stephen, D., Lenke, C., Donner, P., Gärtner, S., Taubert, N., Bausenwein, T., & Stahlschmidt, S. (2024). *The data infrastructure of the German Kompetenznetzwerk Bibliometrie: An enabling intermediary between raw data and analysis*. Zenodo. Retrieved October 28, 2024 from https://doi.org/10.5281/zenodo.13935407

Thelwall, M. (2024a). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*, *9*(2), 1-21.

Thelwall, M. (2024b). *Quantitative Methods in Research Evaluation Citation Indicators, Altmetrics, and Artificial Intelligence* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.00135

Thelwall, M., & Yaghi, A. (2024). *In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2409.16695

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., & Azhar, F. (2023). *Llama: Open and efficient foundation language models* arXiv. https://doi.org/10.48550/arXiv.2302.13971

Vital Jr, A., Silva, F. N., Oliveira Jr, O. N., & Amancio, D. R. (2024). *Predicting citation impact of research papers using GPT and other text embeddings* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2407.19942

Wu, M., Sivertsen, G., Zhang, L., Qi, F., & Zhang, Y. (2024). *Scientific progress or societal progress? A language model-based classification of the aims of the research in scientific publications.* 28th International Conference on Science, Technology and Innovation Indicators (STI2024), Berlin, Germany.

Wu, M., Zhang, Y., Zhang, G., & Lu, J. (2021). Exploring the genetic basis of diseases through a heterogeneous bibliometric network: A methodology and case study. *Technological Forecasting and Social Change*, *164*, 120513.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., & Huang, F. (2024). *Qwen2 technical report* [Preprint]. arXiv. https://doi.org/https://doi.org/10.48550/arXiv.2407.10671

Supplementary Material

Prompt for Task 1

p1 (used for ZS and fine-tuning):

You are an academic expert in the biomedical field, evaluating research articles based on scientific rigor, replicability, data analysis, and study limitations. You will summarize each article as "recommend" or "not recommend" by reading the abstracts.
Reply with 1 for recommending this article and 2 for not recommending it.
Reply with 1 or 2 and nothing else.

p2 (used for ZS and fine-tuning):

You are an academic expert in the biomedical field, evaluating research articles based on scientific rigor, replicability, data analysis, and study limitations. You will summarize each article as "recommend" or "not recommend" by reading the abstracts.

- *Scientific rigor is the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results.*
- *Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.*
- *Data analysis is the practice of working with data to glean useful information, which can then be used to make informed decisions.*
- *Study limitations are the constraints placed on the ability to generalize from the results, to further describe applications to practice, and/or related to the utility of findings that are the result of the ways in which you initially chose to design the study, or the method used to establish internal and external validity or the result of unanticipated challenges that emerged during the study.*

Reply with 1 for recommending this article and 2 for not recommending it.
Reply with 1 or 2 and nothing else.

Additional FS demonstrations:

Here are some examples from human expert recommendations:

Traditional epidural techniques have been ... [Abstract with three or more recommendations]

1 *+ 4 more abstracts in this category*

The nature of "climate change" will differ ... [Abstract with no recommendations]

2 *+ 4 more abstracts in this category*

Prompt for Task 2

p3 (used for rating classification)

You are an academic expert in the biomedical field, evaluating research articles based on scientific rigor, replicability, data analysis, and study limitations. The definitions of the evaluation dimensions are as follows:

- *Scientific rigor is the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results.*
- *Replicability is obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data.*
- *Data analysis is the practice of working with data to glean useful information, which can then be used to make informed decisions.*
- *Study limitations are the constraints placed on the ability to generalize from the results, to further describe applications to practice, and/or related to the utility of findings that are the result of the ways in which you initially chose to design the study, or the method used to establish internal and external validity or the result of unanticipated challenges that emerged during the study.*

You will summarize your rating using 1, 2, or 3, representing "Good," "Very Good," and "Exceptional" quality. Just reply with 1, 2, or 3 and nothing else.

Additional FS demonstrations:

Here are some examples from human expert recommendations:

Significance Fluorescent auxin ... (Abstract with rating of 1)

1 *+ 4 more abstracts in this category*

Several surveys have observed different degrees ... (Abstract with rating of 2)

2 *+ 4 more abstracts in this category*

The long-standing belief that nitrogen-containing ... (Abstract with rating of 3)

3 *+ 4 more abstracts in this category*

https://doi.org/10.51408/issi2025_060

Linking Data Citation to Repository Visibility: An Empirical Study

Fakhri Momeni[1], Janete Saldanha Bach[2], Brigitte Mathiak[3], Peter Mutschke[4]

[1] *fakhri.momeni@gesis.org*, [2]*janete.saldanhabach@gesis.org*, [3]*brigitte.mathiak@gesis.org*, [4]*peter.mutschke@gesis.org*
Knowledge Technologies for the Social Sciences (KTS), GESIS - Leibniz Institute, Unter Sachsenhausen 6-8 50667 Cologne (Germany)

Abstract

In today's data-driven research landscape, dataset visibility and accessibility play a crucial role in advancing scientific knowledge. At the same time, data citation is essential for maintaining academic integrity, acknowledging contributions, validating research outcomes, and fostering scientific reproducibility. As a critical link, it connects scholarly publications with the datasets that drive scientific progress. This study investigates whether repository visibility influences data citation rates. We hypothesize that repositories with higher visibility, as measured by search engine metrics, are associated with increased dataset citations. Using OpenAlex data and repository impact indicators (including the visibility index from Sistrix, the h-index of repositories, and citation metrics such as mean and median citations), we analyze datasets in Social Sciences and Economics to explore their relationship. Our findings suggest that datasets hosted on more visible web domains tend to receive more citations, with a positive correlation observed between web domain visibility and dataset citation counts, particularly for datasets with at least one citation. However, when analyzing domain-level citation metrics, such as the h-index, mean, and median citations, the correlations are inconsistent and weaker. While higher visibility domains tend to host datasets with greater citation impact, the distribution of citations across datasets varies significantly. These results suggest that while visibility plays a role in increasing citation counts, it is not the sole factor influencing dataset citation impact. Other elements, such as dataset quality, research trends, and disciplinary norms, also contribute significantly to citation patterns.

Introduction

In the ever-evolving landscape of scholarly research, the effective citation of data is fundamental in fostering transparency, reproducibility, and the robust advancement of scientific knowledge. As data-intensive research continues to expand, the growing volume and diversity of available datasets make standardized and effective data citation practices increasingly crucial.

This study examines the relationship between repository visibility, repository impact, and dataset citation rates. We define repository visibility through Sistrix's visibility index, which measures how discoverable a repository is via search engines, and repository impact through the h-index of datasets hosted on a domain. The "Joint Declaration of Data Citation Principles" by the Data Citation Synthesis Group (Martone, 2014) represents a key contribution from the FORCE11 community that organized the FAIR data principles. By emphasizing standardized, accessible, and transparent data citation practices, the declaration provides a foundation for this research. Building upon such works, this study aims to offer detailed insights into the interplay between repository visibility and data citation, addressing critical needs in the evolving scholarly landscape.

Lin et al. (2014) initiative underscores the ongoing efforts to establish metrics that quantify data usage and citation, thereby recognizing and attributing credit to data authors for their contributions. Moreover, studies like that of Piwowar and Vision (2013) highlight the potential advantages of open access to data, suggesting a correlation between openly accessible datasets and increased citation rates.

This paper examines the connection between repository visibility (Sistrix index) and repository impact (h-index of hosted datasets) in relation to data citation rates. It also addresses the challenges highlighted by Starr et al. (2015) regarding the accessibility of cited data in scholarly publications. Furthermore, Robinson-Garcia et al. (2017) evaluate DataCite as a bibliometric source, offering insights into the tools available for analyzing data citation trends.

Navigating the general agreement and debates surrounding data publication, as discussed by Kratz & Strasser (2014), this study contributes to the ongoing conversation by quantifying the relationship between repository visibility, repository impact, and dataset citation. By integrating these perspectives, we aim to refine our understanding of how dataset discoverability and repository influence shape citation practices.

We are particularly interested in the following research question: Does the visibility of data repositories, as measured by the Sistrix index, correlate with their citation impact (h-index, mean, and median citations) and the number of citations received by datasets they host?

Related work

Several studies have looked at the challenges of citing data to make it easier to find and access. For example, Krause & Mongeon (2023) examined how datasets are cited in the OpenAlex database. They found interesting patterns in how data creators are connected to the authors who cite their work. Their study revealed trends in citation across different countries and emphasized the importance of open research practices for clear and transparent scholarly work.

Park, You & Wolfram (2018) highlighted the prevalence of informal data citation within scientific publications, particularly in the biological and biomedical sciences, the fields with the most public data sets available documented by the Data Citation Index (DCI). Their study underscored the challenges in adequately acknowledging and documenting data contributions alongside formal data citation practices, emphasizing the need for streamlined citation methodologies to encompass informal data attributions effectively.

Addressing the imperative for FAIR principles (findability, accessibility, interoperability, reusability) in biomedical datasets, Tsueng et al. (2023) highlighted the challenges associated with achieving FAIRness. Their findings emphasized the absence of a unified metadata standard among repositories housing these datasets, hindering the discoverability and accessibility of datasets on major platforms such as Google Dataset Search. This underscores the importance of metadata standardization in improving dataset accessibility and visibility.

Innovative approaches proposed by Bach, Klas & Mutschke, (2022) introduced an infrastructure to assign Persistent Identifiers (PIDs) to dataset elements (i.e.,

variables) within Social Sciences datasets. This novel framework tackled data citation and reuse obstacles by offering a structured method to reference specific elements within data files, thus facilitating retrieval with requisite metadata. This approach catered to both machine-actionable and human-accessible needs, significantly improving data citation practices in the Social Sciences fields.

Groth et al. (2020) emphasized the pivotal role of data citation and its underlying infrastructures, particularly associated metadata, in enabling FAIR data reuse. This paper underscored the importance of data citation in rendering datasets findable and accessible. It advocated for consistently implementing and supporting machine-readable metadata to address challenges hindering maximal and appropriate reuse of existing datasets.

Onyancha (2016) found a strong correlation between data citation and article citation (correlation coefficient of 0.68) as well as between data citation and the h-index of journals (correlation coefficient of 0.71). These findings highlight the interconnectedness between data citation practices and the scholarly impact of articles within their respective journals.

While these studies have significantly contributed to understanding data citation dynamics and enhancing data accessibility, there remains a notable research gap concerning the direct impact of repository visibility and data access status on data citation rates. This study aims to investigate the intricate relationship between data citation, repository findability, and data access in the Social Sciences and Economics fields, aiming to elucidate how these factors influence citation rates. By addressing these factors, the study seeks to benefit researchers by promoting proper data citation practices, which can enhance research visibility, foster academic credibility, and ensure acknowledgment of contributors' work. Additionally, improving data citation practices benefits the broader research community by facilitating transparency, reproducibility, and collaboration. This makes data more discoverable and accessible, directly appealing to researchers as key stakeholders who rely on well-documented and accessible datasets.

The central research question involves uncovering how much repository visibility and data access status correlate with data citation rates, thereby highlighting the critical need to engage researchers and repository managers in adopting practices that improve data sharing and citation.

Data and Methods

Data Source and Dataset Selection

We utilized the 2023 version of the OpenAlex database, maintained by the German Competence Centre for Bibliometrics, to gather bibliometric information on published datasets. It offers openly available metadata on scholarly works, including publications, datasets, authors, institutions, journals, and research topics, facilitating comprehensive analyses of research outputs and their impact. We identified datasets by filtering those where the *'item_type'* value was *'dataset'* within the table containing published items in the database. From the expansive collection of datasets covering

284 subject categories, our focus centered on datasets associated with the following fields: Economic[1] and Social Sciences[2].
Our dataset compilation comprised 401,659 datasets in the 'Social Sciences' category and 78,267 datasets in the 'Economics' category.
There is an overlap involving 37,160 datasets in the 'Social Sciences' and 'Economics' categories.
A notable observation emerged when analyzing the citation rates of datasets within these fields in OpenAlex. Surprisingly, our findings revealed that most datasets in these fields had no citations recorded in OpenAlex. Specifically, within the 'Social Sciences' category, a staggering 99% of datasets had no recorded citations, while in the 'Economics' category, 98% of datasets remained uncited. This remarkable prevalence of uncited datasets raises critical questions about the relevance of these datasets, the extent to which they are reused in scholarly research, and whether they are being utilized without proper citation, highlighting a significant challenge in promoting data attribution and recognition within academic work.
We extracted digital object identifiers (DOIs) from datasets listed in OpenAlex and determined their primary web domains using a Python script.

Key Measures: h-Index and Visibility Index

To explore the factors associated with dataset citation rates, we focus on two key measures: the visibility index and h-index. These metrics were selected for their complementary roles in capturing the scholarly impact and visibility of web domains hosting datasets. By investigating these measures, we aim to uncover patterns that might inform strategies for improving dataset attribution and visibility.
The visibility index, obtained from Sistrix[3], is widely used in digital marketing and search engine optimization. It measures how visible a web domain is in search engine results, offering insight into its discoverability. It is done in three steps: collection of data, weighting of data and summation of the values for the visibility index. A detailed explanation of this methodology is available in Sistrix's official documentation[4]. In our study, we compute a web domain's visibility index by averaging its scores across 30 countries, as visibility plays a crucial role in dataset findability and potential citation impact. To retrieve visibility index values, we

[1]It includes Development Economics, Economic Geography, Financial Economics, Law and Economics, Economic History, Economic System, Environmental Economics, Keynesian Economics, Econometrics, Labour Economics, Natural Resource Economics, Socioeconomics, Neoclassical Economics, Demographic Economics, Mathematical Economics, Welfare Economics, Political Economy, Economic Policy, Economic Growth, Market Economy, Development Economics, Public Economics, Macroeconomics, Monetary Economics, International Economics, Agricultural Economics, Economic Geography, Positive Economics, Classical Economics, Economy, Finance, Financial System, and Business Administration

[2]Encompassing disciplines such as Social Science, Archaeology, Anthropology, Developmental Psychology, Law, Library Science, Linguistics, Political Economy, Communication, Demography, Gender Studies, Public Relations, Public Administration, Social Psychology, Socioeconomics, Pedagogy, Management Science, and Management

[3] https://www.sistrix.com/api/domain/domain-visibilityindex/

[4] https://www.sistrix.com/visibility-index/calculation

utilized the *Sistrix API*[5] and developed a script that queries the visibility scores for each domain across multiple country-specific search engine indexes. This automated approach ensures consistency and scalability in our data collection process. The full code is available at: GitHub Repository[6].

Traditionally, the h-index has been used to measure the impact of authors or journals based on citation data (Mester, 2016). In this study, we extend its application to measure the impact of web domains hosting datasets. This novel adaptation evaluates how influential a domain is in facilitating impactful research through its datasets.

The h-index of a web domain is calculated by ranking its datasets in descending order based on the number of citations and finding the point where the rank equals or exceeds the number of citations. Mathematically:

$$h = \max\{h' : h' \leq c_{h'}\}$$

where h' is the rank of a dataset, and $c_{h'}$ is the number of citations of the dataset ranked h'. For example, a domain with an h-index of 20 has at least 20 datasets, each cited 20 or more times, while all other datasets have fewer than 20 citations.

This adaptation leverages the h-index's ability to account for both the quantity (number of datasets) and the quality (number of citations) of datasets hosted on a web domain, providing a balanced metric of impact. Similar to its application in author-level metrics, the h-index for web domains highlights the extent of reuse and scholarly influence of the datasets they host.

Data Filtering and Analysis

To focus on datasets that reflect current trends and practices in data citation and visibility, we excluded datasets published before 2016. This decision ensures that the analyzed data remains relevant and aligns with the dynamic nature of the visibility index and the evolving landscape of research dissemination. Consequently, we applied a filtering process to identify newer datasets published from 2016 onward. This systematic approach led us to analyze a total of 155,564 datasets in Social Sciences and 37,621 in Economics, with an overlap of 18,998 datasets between the two fields.

The number of datasets associated with each web domain, their respective average visibility index (have been acquired from Sistrix in August and September 2023), and the computed h-index for each web domain (based on the citation counts of datasets hosted on the web domain) can be accessed on GitHub.

We computed the mean and standard deviation for the h-index and visibility index to offer statistical insights into the data's distribution and central tendencies. We also computed the mean and standard deviation of normalized citation[7] for the datasets published during 2016 and 2023 (155,564 in Social Sciences and 37,621 in

[5] https://www.sistrix.com/api/domain/domain-visibilityindex/

[6] https://github.com/momenifi/Dataset_finability/blob/main/2visibility_domains.py

[7] by dividing the number of citations by the age of publication (age of publication is equal to 2023 - publication year)

Economics). **Table *1*** displays the mean and standard deviation. The notably high standard deviation observed in both the Social Sciences and Economics categories indicates considerable variability within the dataset.

Correlation Analysis

Given the high variability in citation counts and other metrics, we employed Spearman's coefficient to determine the correlation between these variables. Spearman correlation, as a non-parametric measure, evaluates the strength and direction of monotonic relationships rather than relying on the actual values, rendering it less sensitive to extreme values or outliers. Given its resilience to extreme values, Spearman's correlation is a more suitable metric for depicting the monotonic relationship between these variables.

Table 1. Mean and Standard Deviation (SD) for the h-Index and the Visibility Index of web domains, and the number of citations received by datasets.

Variable	*Mean (Social Sciences)*	*SD (Social Sciences)*	*Mean (Economics)*	*SD (Economics)*
h-index	0.95	5.1	0.89	4.84
Visibility Index	0.22	3.03	0.32	3.9
Normalized citation	0.08	2.07	0.14	1.76

Results

We investigated how web domain visibility indices relate to the citation impact of their respective datasets. To gauge the citation impact within each web domain, we calculated the eight-year h-index for each web domain using published datasets from 2016 to 2023, focusing on the *Social Sciences* and *Economics* categories. The eight-year h-index for a web domain, denoted as *h*, indicates that the web domain *d* should have at least *h* datasets published between 2016 and 2023, with each dataset receiving a minimum of *h* citations.

Tables ***A.1, A.2, A.3, A.4, A.5***, and ***A.6*** present the top ten web domains based on the number of datasets, web domains' h-index, and visibility index. ***Table A.5*** and ***Table A.6*** illustrate the ten most influential web domains (in terms of the h-index) under which datasets have been registered. Notably, not all these web domains are traditional data repositories. Brill-online is a collection of annotated historical texts, some of which are highly relevant to Social Sciences and Economics theory, and Psycnet is mainly a journal for Psychology, which also stores other entity types, such as instruments, which are often used in questionnaires used to conduct surveys. While these instruments are categorized as datasets in the metadata, they are not what one typically envisions when talking about datasets, just like the historical texts. However, we would argue they are in the same category as datasets, in that they are resources produced by researchers for other researchers to enable or facilitate research, and they follow the same rules. Additionally, both annotated texts and

instruments are both commonly referred to as "data" in their respective home disciplines: literary studies and psychology. We see this as an example of how research data infrastructures designed for one discipline elevate research in other disciplines as well.

There are some systematic biases in play as well. ICPSR is one of the largest data providers for quantitative data. However, there are only a handful of citations registered in the dataset. We know that citation of data is often done informally (Boland et al., 2012), therefore we assume that the use of these datasets is strongly underreported.

The **Table *2*** and **Table *3*** summarize the relationship between the visibility index of web domains and their citation impact in the fields of *Social Sciences* and *Economics*. These tables present the Spearman correlation coefficients for different h-index threshold levels, illustrating how the strength of association between visibility and impact changes across subsets of web domains. In addition to the h-index, the analysis also includes mean and median citations, providing a more detailed view of how web visibility relates to different aspects of citation impact.

The results indicate a positive correlation between visibility and the h-index across all thresholds, with stronger associations observed at higher h-index levels. This suggests that domains with greater web visibility tend to be linked to more impactful research, particularly among the most highly cited domains. However, the correlation between visibility and citation counts (both mean and median) is weaker and inconsistent. In Social Sciences, mean citations show a slight positive correlation at the broadest threshold but turn negative at higher h-index thresholds, while median citations exhibit no clear pattern. In Economics, visibility has little to no correlation with citation counts, suggesting that highly cited research does not necessarily originate from highly visible domains.

These findings reinforce the idea that web visibility is associated with domain-level research impact, as measured by the h-index, but does not directly translate into higher citation counts. While visibility may enhance discoverability, other factors (such as dataset quality, research trends, and disciplinary practices) play a crucial role in shaping citation impact.

Table 2. Spearman Correlation between h-index of web domains, dataset's citation metrics (mean and median citation) in *Social Sciences* and visibility index of web domain captured by Sistrix. The table presents correlation values along with their respective p-values in parentheses.

h-index Threshold	*Number of Web Domains*	*h-index Correlation*	*Mean Citations Correlation*	*Median Citations correlation*
All web domains	389	0.14 (0.005)		-0.043 (0.393)
Web domains with h-index > 0	144	0.23 (0.004)	-0.112 (0.181)	-0.147 (0.078)

Web domains with h-index > 1	55	0.37 (0.005)	-0.002 (0.986)	-0.004(0.975)

Table 3. Spearman Correlation between h-index of web domains, dataset's citation metrics (mean and median citation) in *Economics* and visibility index of web domain captured by Sistrix. The table presents correlation values along with their respective p-values in parentheses.

h-index Threshold	*Number of Web Domains*	*h-index Correlation*	*Mean Citations Correlation*	*Median Citations correlation*
All web domains	225	0.14 (0.040)	0.097 (0.148)	0.008 (0.908)
Web domains with h-index > 0	84	0.31 (0.004)	0.0001 (0.999)	-0.055 (0.622)
Web domains with h-index > 1	25	0.47 (0.017)	-0.028 (0.895)	-0.052 (0.805)

Table *4* and **Table *5***[8] provide additional insights by examining correlations at the dataset level. The results show that visibility correlates more strongly with whether a dataset receives at least one citation rather than with citation counts beyond the first citation. This aligns with prior research emphasizing that literature plays a key role in helping researchers discover datasets (Gregory et al., 2020). The correlation between visibility and first citation suggests that repository visibility plays an essential role in the initial citation of datasets, but other factors such as dataset quality, disciplinary norms, and research trends may be more influential in determining long-term citation impact.

Overall, the findings suggest that web domain visibility is associated with domain-level research impact, as measured by the h-index, but does not directly translate into higher citation counts at the dataset level. While visibility may enhance discoverability, other factors play a crucial role in shaping citation impact. Furthermore, it is important to acknowledge that correlation does not imply causation. While the associations observed in this study suggest a relationship between visibility and impact, further investigation is required to understand the causal mechanisms underlying these patterns.

Many web domains lacked citations for their datasets, likely due to the widespread practice of informal data citation for data sharing and reuse in research papers, as highlighted by Park, You & Wolfram (2018). Their study in the biological/biomedical sciences field highlighted the prominence of informal data citations within the main text of articles, contrasting with the less frequent occurrence of formal data citations within references. Considering the significant

[8] The code available here: https://github.com/momenifi/Dataset_finability/blob/main/3correlation_analysis.py

number of datasets and web domains lacking citations, we performed correlation calculations under different conditions: first, without any filtering; second, by exclusively including web domains with an h-index greater than 0; and third, by also encompassing web domains with an h-index greater than one.

As a whole, the h-index of web domains for data repositories, as defined in this study, serves as a plausible indicator of the expected citation rate for datasets in the Social Sciences and Economics, including those not traditionally classified under these fields. The highly cited datasets tend to be clustered together. Most influential websites are also well-known in the community and offer plausible and useful contributions. Additionally, we have an indication that these findings likely extend to other forms of scholarly artifacts that are not always explicitly referred to as data. However, this aspect requires further investigation to be fully understood.

Table 4. Correlation between datasets' number of received citations in *Social Sciences* and visibility index of web domain captured by Sistrix.

Threshold	*Number of Datasets*	*Spearman Correlation (P-Value)*
All datasets	146730	0.14 (0.0)
Datasets with number of citations > 0	7183	0.31 (0.0)
Datasets with number of citations > 1	1854	0.01 (0.633)

Table 5. Correlation between datasets' number of received citations in Economics and visibility index of web domain captured by Sistrix.

Threshold	*Number of Datasets*	*Spearman Correlation (P-Value)*
All datasets	34969	0.15 (0.0)
Datasets with number of citations > 0	2932	0.37 (0.0)
Datasets with number of citations > 1	946	-0.02 (0.555)

Conclusion and Discussion

The fact that 99% of Social Sciences datasets and 98% of Economics datasets have no citations is a big challenge. Onyancha (2016) underscores the challenges in data citation compared to research articles, particularly in sub-Saharan Africa (SSA), where data citation rates are notably lower. It makes us wonder why these datasets aren’t getting cited much. To address these challenges, increasing awareness among researchers about proper data citation is crucial. Establishing clear citation standards and providing incentives for proper attribution could encourage better citation practices. Additionally, improving metadata quality and repository infrastructure will enhance the discoverability of datasets, making them more accessible for researchers. Some researchers might be sharing data informally. Researchers may

reuse data informally, without following the usual ways of citing it, such as the Joint Declaration of Data Citation Principles (Martone, 2014).
Also, some repositories do not make datasets easily findable, which impacts their citation rates. To fix this, they need to improve metadata standards to enhance dataset discoverability. We could make clear rules for citing data and give rewards to encourage people to do it properly, raising awareness among researchers in Social Sciences and Economics about proper data citation. Also, encourage repositories to optimize visibility through structured metadata and persistent identifiers (PIDs). we should make it easier to find these datasets by improving the information about them in the places they're stored.
Our correlation analysis suggests that datasets hosted on more visible web domains tend to receive more citations. At the dataset level, we find a positive correlation between dataset citation counts and web domain visibility, particularly for datasets with at least one citation. However, when analyzing domain-level citation metrics, such as the h-index, mean citation, and median citation, the correlations are less consistent. While higher visibility domains tend to host datasets with greater overall citation impact, the distribution of citations across datasets varies widely. Importantly, these correlations do not imply a causal relationship—higher domain visibility does not necessarily predict higher dataset citation rates. Instead, other factors, including dataset quality, repository policies, and researcher behaviors, play a role in shaping both visibility and impact.
To address these challenges and improve dataset findability and citation rates, repositories should consider implementing structured metadata, persistent identifiers, and FAIR Signposting. FAIR Signposting, a lightweight mechanism using standardized HTTP link headers, can enhance dataset discovery and usability by guiding researchers and automated tools toward relevant metadata, licensing information, and dataset relationships (Wilkinson et al., 2022). Integrating FAIR Signposting into repositories could:

- Make datasets easier to find by enabling automated tools and search engines to retrieve structured metadata.
- Improve citation tracking by linking datasets directly to related publications and persistent identifiers (PIDs).
- Facilitate better metadata interoperability, ensuring datasets are consistently indexed and referenced across different platforms.

Overall, while our findings suggest a relationship between repository visibility and citation impact, significant efforts are needed to improve dataset discoverability, formalize citation practices, and encourage researchers to attribute data properly. Addressing these challenges will be crucial in ensuring that datasets in Social Sciences and Economics gain the recognition and reuse they deserve.

Limitation

We acknowledge several limitations in this study. First, the challenge of studying data citation is evident, as many datasets in OpenAlex lack citations. It remains unclear whether these datasets receive fewer citations than published papers or if authors often rely on informal citations for data sharing and reuse (Park, You & Wolfram, 2018). This results in a skewed distribution of citation data, which may limit the accuracy of correlation analysis and lead to an underestimation of the impact of repository visibility on citation rates. The inaccessibility of informal citations may lead to incomplete correlations, limited insights, and potential underestimation of the impact on repository discoverability.

Additionally, while our analysis encompassed a wide array of datasets from OpenAlex, it's essential to note that our examination did not specifically quantify or track the prevalence of atypical datasets, such as online presentations in video format, within the dataset.

While these atypical datasets were part of our dataset, their specific prevalence or impact wasn't quantified or tracked within our analysis. Future research might benefit from quantifying these unconventional datasets' prevalence and impact on data citation practices.

Data availability

Data and code are accessible via the following link: https://github.com/momenifi/methodHub/blob/main/dataset_findability/

References

Bach, J. S., Klas, C.-P., and Mutschke, P. (2022). The hurdles of current data citation practices and the adding-value of providing pids below study level. *In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5.

Boland, K., Ritze, D., Eckert, K., and Mathiak, B. (2012). Identifying references to datasets in

publications. *In Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings 2*, pages 150–161. Springer.

Gregory, K., Groth, P., Scharnhorst, A., and Wyatt, S. (2020). Lost or found? discovering data needed for research. Harvard Data Science Review, v. 2, no. 2

Groth, P., Cousijn, H., Clark, T., and Goble, C. (2020). Fair data reuse–the path through data citation. *Data Intelligence*, 2(1-2):78–86.

Kratz, J. and Strasser, C. (2014). Data publication consensus and controversies. *F1000Research*, 3.

Krause, G. and Mongeon, P. (2023). Measuring data re-use through dataset citations in OpenAlex. *In 27th International Conference on Science, Technology and Innovation Indicators (STI 2023). International Conference on Science, Technology and Innovation Indicators.*

Lin, J., Cruse, P., Fenner, M., and Strasser, C. (2014). Making data count: A data metrics pilot

project.

Martone, M., editor (2014). Joint Declaration of Data Citation Principles. *FORCE11, San Diego, CA.*

Mester, G. (2016). Rankings scientists, journals and countries using h-index. *Interdisciplinary Description of Complex Systems*: INDECS, 14(1):1–9.

Onyancha, O. B. (2016). Open research data in sub-saharan africa: a bibliometric study using the data citation index. *Publishing Research Quarterly*, 32(3):227–246.

Park, H., You, S., and Wolfram, D. (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, 69(11):1346–1354.

Piwowar, H. A. and Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, 1:e175.

Robinson-Garcia, N., Mongeon, P., Jeng, W., and Costas, R. (2017). Datacite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, 11(3):841–854.

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., Haak, L. L., Haendel, M., Herman, I., Hodson, S., et al. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, 1:e1.

Tsueng, G., Cano, M. A. A., Bento, J., Czech, C., Kang, M., Pache, L., Rasmussen, L. V., Savidge, T. C., Starren, J., Wu, Q., et al. (2023). Developing a standardized but extendable framework to increase the findability of infectious disease datasets. *Scientific Data*, 10(1):99.

Wilkinson, M. D., Sansone, S.-A., Marjan, G., Nordling, J., Dennis, R., and Hecker, D. (2022). Fair as-

sessment tools: towards an “apples to apples” comparisons. *URL: https://doi. org/10.5281/zenodo,*

7463421.

Appendix

Appendix A

This appendix presents the top web domains based on various criteria. Tables are provided to showcase the top ten web domains with the highest number of DOIs, mean visibility index, and h-index in both the Economics and Social Sciences datasets.

Table A.1. Top ten web domains with highest DOI numbers (Economics).

Domain	*DOIs Number*	*Mean Visibility Index*	*h-Index*
https://referenceworks.brillonline.com/	9128	0.122337	74
https://primarysources.brillonline.com/	8941	0.000383	11
https://www.socialscienceregistry.org/	5676	0.000933	9
https://psycnet.apa.org:443/	1649	0.182410	8
https://connect.h1.co/	1639	0.000000	2
https://www.oecd-ilibrary.org/	1594	0.285173	12
https://www.healthaffairs.org/	1301	0.041977	3
https://www.openicpsr.org/	1212	0.001127	1
https://www.iucnredlist.org/	858	0.073103	2
https://www.degruyter.com/	576	2.044188	5

Table A.2. Top ten web domains with highest DOIs (Social Sciences).

Domain	*DOIs Number*	*Mean Visibility Index*	*h-Index*
https://primarysources.brillonline.com/	45967	0.000383	18
https://referenceworks.brillonline.com/	26343	0.122337	98
https://scholarlyeditions.brill.com/	18509	0.000000	1
https://psycnet.apa.org:443/	16820	0.182410	23
https://connect.h1.co/	7599	0.000000	9
https://www.socialscienceregistry.org/	7113	0.000933	6
https://connect.liblynx.com/	4715	0.000000	0
https://www.degruyter.com/	2649	2.044188	7
https://www.healthaffairs.org/	1961	0.041977	3
https://doi.pangaea.de/	1800	0.001448	4

Table A.3. Top ten web domains with highest visibility index (Economics).

Domain	*DOIs Number*	*Mean Visibility Index*	*h-Index*
https://www.youtube.com/	2	61.709560	0
https://www.researchgate.net/	8	4.404945	0
https://onlinelibrary.wiley.com/	23	2.051243	1
https://www.degruyter.com/	576	2.044188	5
https://www.fao.org/	216	1.310798	1
https://www.cambridge.org/	3	0.950633	0
https://www.fs.usda.gov/	20	0.784208	1
https://www.science.org/	166	0.770167	4
https://www.ebi.ac.uk/	2	0.570627	0
https://www.erudit.org/	1	0.448252	0

Table A.4. Top ten web domains with highest visibility index (Social Sciences).

Domain	*DOIs Number*	*Mean Visibility Index*	*h-Index*	
https://www.youtube.com/	4	61.709560	0	
https://link.springer.com/	5	6.462607	0	
https://www.researchgate.net/	28	4.404945	0	
https://www.jstor.org/	1	3.635153	0	
https://onlinelibrary.wiley.com/	86	2.051243	1	
https://www.degruyter.com/	2649	2.044188	7	
https://www.fao.org/	271	1.310798	1	
https://www.cambridge.org/	2	0.950633	0	
https://www.fs.usda.gov/	370	0.784208	6	
https://www.science.org/	1241	0.770167	10	

Table A.5. Top ten web domains with highest h-index (Economics).

Domain	*DOIs Number*	*Mean Visibility Index*	*h-Index*
https://referenceworks.brillonline.com/	9128	0.122337	74
https://www.oecd-ilibrary.org/	1594	0.285173	12
https://primarysources.brillonline.com/	8941	0.000383	11
https://www.socialscienceregistry.org/	5676	0.000933	9
https://psycnet.apa.org:443/	1649	0.182410	8
https://www.degruyter.com/	576	2.044188	5
https://www.science.org/	166	0.770167	4
https://www.healthaffairs.org/	1301	0.041977	3
https://www.authorea.com/	372	0.007120	3
https://doi.pangaea.de/	107	0.001448	3

Table A.6. Top ten web domains with highest h-index (Social Sciences).

Domain	*DOIs Number*	*Mean Visibility Index*	*h-Index*
https://referenceworks.brillonline.com/	26343	0.122337	98
https://psycnet.apa.org:443/	16820	0.182410	23
https://primarysources.brillonline.com/	45967	0.000383	18
https://www.oecd-ilibrary.org/	1451	0.285173	11
https://www.science.org/	1241	0.770167	10
https://connect.h1.co/	7599	0.000000	9
https://www.degruyter.com/	2649	2.044188	7
https://www.socialscienceregistry.org/	7113	0.000933	6
https://oxfordbibliographies.com/	418	0.000000	6
https://www.fs.usda.gov/	370	0.784208	6

https://doi.org/10.51408/issi2025_068

Linking Research Publications to SDGs: Exploring the SDG Mapping in Web of Science, Scopus and OpenAlex

Prashasti Singh[1], Vivek Kumar Singh[2], Anurag Kanaujia[3], Abhirup Nandy[4]

[1]*prashasti.singh8@gmail.com*
Department of Computer Science, Shri Ram College of Commerce, University of Delhi, Delhi-110007 (India)

[2]*vivekks12@gmail.com*
Department of Computer Science, University of Delhi, Delhi-110007 (India)
NITI Aayog, Government of India, Delhi-10007 (India)

[3] *anuragkanaujia01@gmail.com*
Delhi School of Analytics, University of Delhi, Delhi-110007 (India)

[4]*abhirupnandy.online@gmail.com*
Institute of Informatics and Communication, University of Delhi, Delhi-110067 (India)

Abstract

The Sustainable Development Goals (SDGs) are 17 global objectives proposed by the United Nations to create a better and more sustainable future by 2030. Since the adoption of SDGs in 2015, several missions and programmes have been initiated across different countries towards achieving the relevant targets under SDGs. The advancements in science and technology research and development is believed to play a crucial role in achieving these targets. Motivated by the role of scientific outcomes in SDGs, various scholarly databases have started to map the indexed research publications to one or more of the SDGs. A host of approaches (employing keyword based, machine learning, manual curation etc.) have been used to link and map research publications under the SDGs. Some initial studies have shown that the mapping of publications in SDGs vary significantly across different databases. However, the classification accuracy, thematic focus, practical applicability, and impact of these classification approaches have not been studied well. Therefore, this work attempts to make a deeper exploration of the SDG mapping in three major scholarly databases- Web of Science, Scopus and OpenAlex and provide useful insights. For this purpose, a large-scale data sample of publications for the year 2023 obtained from these three databases are analysed on different aspects. Results suggest that not only the three databases vary significantly in terms of their individual SDG mapping, but there are also significant differences in the SDG-wise distribution and interlinkages across different SDGs. A divergence score to measure divergence of classification across the three databases is defined and computed. Finally, the probable reasons and factors that may be resulting in the variations in SDG mappings across the three databases are explored and discussed.

Introduction

The Sustainable Development Goals (SDGs) are 17 global objectives established by the United Nations in 2015, aiming to create a better and more sustainable future by 2030. These goals address critical challenges like no poverty (SDG 01), zero hunger (SDG 02), good health & well-being (SDG 03), quality education (SDG 04), gender equality (SDG 05), clean water and sanitation (SDG 06), affordable and clean energy (SDG 07), decent work and economic growth (SDG 08), industry, innovation and infrastructure (SDG 09), reduced inequalities (SDG 10), sustainable cities and

communities (SDG 11), responsible consumption and production (SDG 12), climate action (SDG 13), life below water (SDG 14), life on land (SDG 15), peace, justice and strong institutions (SDG 16), and partnerships for the goals (SDG 17). Since their adoption in 2015, they have made a universal call to action and motivated significant research and development of novel technologies (United Nations, 2015). Some of the targets are interlinked and require global cooperation and partnership to achieve, promoting inclusive and sustainable development for all (Sachs, 2012).

The advancements in science and technology research and development are believed to play a crucial role in achieving several targets under SDGs (IISD, 2021; Singh et al., 2024). Research provides evidence-based insights, fosters innovation, and enables the development of sustainable and scalable interventions. It also promotes collaboration among governments, academia, the private sector, and civil society. The role of technology in achievement of SDGs has been studied and underlined (IISD, 2021). Motivated by the role of scientific outcomes in SDGs, various scholarly databases have started to map the indexed research publications to one or more of the SDGs. Scholarly databases are the primary source of providing metadata of scientific outcomes (such as publications and patents), to help identify outcomes that are related to SDGs. A host of approaches (employing keyword based, machine learning, manual curation etc.) have been used to link and map research publications under different SDGs.

Some initial studies have shown that the mapping of publications in SDGs vary significantly across different databases. For example, Armitage, Lorenz, Mikki, (2020) explored the SDG mapping of scholarly publications to identify whether independent bibliometric approaches yield the same results. Another study (Purnell, 2022) compared different methods of identifying publications related to the Sustainable Development Goal 13 on Climate Action and found significant variations across them. Some other studies (such as Hajikhani, & Suominen, 2022; Kashnitsky et al., 2024) also tried to explore SDG mapping of STI outputs in various respects. However, the classification accuracy, thematic focus, interlinkages, and divergence of classification across different scholarly databases have not been studied well. Therefore, this work attempts to make a deeper exploration of the SDG mapping in three major scholarly databases- Web of Science, Scopus and OpenAlex and provide useful insights. This study uses a large-scale data sample of publications for the year 2023 obtained from these three databases and attempts to characterize the SDG mappings in the three databases, not only in terms of variations but also in respect of interlinkages and classification divergence.

Related Work

The research addressing challenges of sustainable development is fairly distributed across the different subject areas. As the impact of challenges covered under the SDGs (e.g., poverty, healthcare, water & sanitation, gender equality and climate change) is becoming more obvious, more research attempting to address the related challenges has been undertaken across the world (Moyer & Hedden, 2020). However, like any other collective exercise, in absence of proper tracing and recapitulation, the cumulative impact from these exercises may end up being a zero-

sum game. In academic research, the estimations have focused on the engagement of researchers and academic staff in universities across the world as key players in promoting SDGs. Studies have provided useful methodologies and insights for analytical exercises in assessment of research in MDGs and SDGs over the years. Keyword analysis is a widely used method for research topics and knowledge mapping. This has been utilized by authors to identify and study the research in SDGs (Armitage, Lorenz, & Mikki, 2020 and Bautista-Puig *et al.,* 2020).

Studies on publication metadata for global research output have explored major trends in research publications. Between the period of 2015 to 2019, the United States, United Kingdom and China are among the top three active countries for research in the different SDGs. SDG 17 i.e., partnerships for SDGs, has the most research publications associated with it, followed by SDG 13, i.e., climate action. Other SDGs with high research activity include SDG 12 (responsible consumption and production), SDG 15 (life on land), SDG 3 (good health and well-being), and SDG 1 (no poverty) (Sweileh, 2020). Co-citation occurrences showed that SDG 11 and SDG 3 are closely related and are frequently referred together in research publications (Bautista-Puig *et al.,* 2020). Most of the research corresponding to SDGs is published from research areas of Life sciences & Biomedicine, and Social Sciences (Meschede, 2020). These studies have started the important process of exploration of research trends in SDGs, and cover only a bird's eye view of the global research trends.

The study by Armitage, Lorenz, Mikki, (2020) explored the SDG mapping of scholarly publications to identify whether independent bibliometric approaches get the same results. Purnell (2022) compared four methods of identifying research publications related to United Nations Sustainable Development 13. These and some other studies identified variations across different scholarly databases. However, the classification accuracy, thematic focus, SDG classification interlinkages, and divergence of classification across major scholarly databases is yet to be suitably explored and analysed. This work attempts to address some of these gaps and provide useful insights about the SDG mapping of research publications across the three major databases.

Data & Method

The study utilized a large-sized data sample of research publications obtained from the three databases- Web of Science (WoS), Scopus, and OpenAlex through their user interfaces (UI). It focused on publication records from the country "India" for the year 2023, limited to document types "article" and "review,". India is one of the major advocates of Sustainable development and has significant focus on research in SDGs (Singh *et al.,* 2022). Further, the practical considerations of availability of data access has also motivated us to use this data as a sample for analysis. The data related to each of the SDGs was downloaded from all the three databases using appropriate search queries or filters provided by the databases. The choice of databases was motivated by the facts that Scopus and Web of Science are among the most popular and reliable databases for scholarly data. Open Alex, has emerged as a large repository of scholarly data providing openly accessible data for Scientometric

application and other research purposes. In addition, these three databases provide SDG classification of records which was an essential consideration. While WoS and OpenAlex offer SDG-based filters in their UI, Scopus provides pre-formulated queries[1] for each SDG enabling data downloads with additional filters. The downloaded search results were pre-processed to eliminate NaN values and duplicate DOIs across all three databases. After this step, the unique records mapped with an SDG were 79,099 in WoS, 69,834 in Scopus and 214,873 in OpenAlex. The processed records in each database were then analysed to understand different patterns. The process of computing analytical results is detailed below.

The SDG wise distribution of publication records in each database is determined**.** Thereafter, the overlapping and unique DOIs for all SDGs across the three databases was identified. The SDG wise mapping of common DOIs across each pair of databases is computed, i.e. for WoS-Scopus, Scopus-OpenAlex and WoS-OpenAlex. The next step was to identify the interlinkages of classification. For this purpose, for a given database, each publication record was scanned to see in which other SDGs it is classified. In this way, all the interlinkages of SDG classification for publications from a given database were done. Similar process was followed for the other databases. The linkages were plotted on a network diagram. The edge weights are proportional to the number of publication records classified in the two connected SDGs together. Absence of any edge in a graph indicates that the two SDGs have no commonly tagged records.

The next computation involved calculating the divergence in SDG mapping of the three databases. Here, first the proportionate share of publication records classified in only one SDG, or two SDGs, or three, or more, is determined for all the three databases. Thereafter a numerical value of Divergence score is proposed and computed. The score can be defined as follows: Given a database, having x number of records mapped to a given SDG, the sum total of all other SDGs in which these x publication records are mapped is used to calculate a Divergence value for the given SDG. Similar values are computed for all other SDGs. Then, all such values are aggregated to compute the overall divergence score of that database. The divergence score for other databases can be computed in a similar way.

$$Divergence\ Value\ (SDGi) = \frac{\sum_{j=1}^{17} x_{i \to j}}{TP_{SDGi}} \quad \ldots(1)$$

$$Divergence\ Score = \frac{\sum_{i=1}^{17} SDGi}{No.of\ SDGs} \quad \ldots(2)$$

where i = SDG number for which divergence is to be calculated, and $x_{i \to j}$ = number of records tagged under SDGi and SDGj, and, TP_{SDGi} = total number of records in the particular SDGi in the given database).

[1]Elsevier SDG Mapping: https://elsevier.digitalcommonsdata.com/datasets/y2zyy9vwzy/1

Results

Variation in Publication Volume mapped with SDGs across the three databases

In terms of total publication count; WoS reported 124,266 research publications for India during 2023, Scopus 194,965 and OpenAlex 340,900 for article and review types. After pre-preprocessing the SDG-wise downloaded data for duplicate and NaN DOIs; 79,099 DOIs were found for WoS, 69,834 for Scopus and 214,873 for OpenAlex. Thus, WoS and OpenAlex were seen to comprise a comparable share of SDG publications in total publications of India amounting to approximately 63% while Scopus reported 35.82% of SDG publications in total publications for India (**Table 1**).

Table 1. Total Publications and Publications tagged with a SDG.

Database	TP	Publications tagged with a SDG	%age
WoS	1,24,266	79,099	63.65
Scopus	1,94,965	69,834	35.82
OpenAlex	3,40,900	2,14,873	63.03

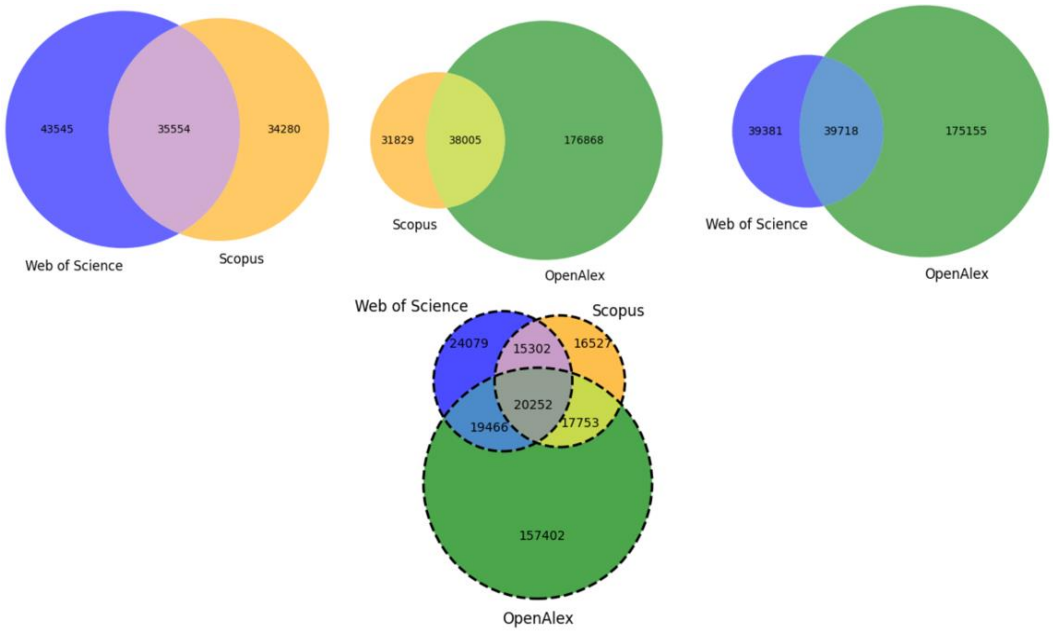


Figure 1. Overlaps in SDG tagged data across the three databases.

Thereafter, the pairwise overlaps across the databases as well as overall overlap across all the three databases together was also computed (**Figure 1**). The relevant numbers can be summarised as follows:

Web of Science- total pre-processed DOIs = 79,099

- Overlap with Scopus- 35,554 (44.95% of WoS), 43,545 DOIs are non-overlapping (55.05% of WoS)

- Overlap with OpenAlex- 39,718 (50.21% of WoS), 39,381 DOIs are non-overlapping (49.79% of WoS)
- Unique DOIs in WoS- 24,079 (30.44% of WoS)

Scopus- total pre-processed DOIs = 69,834

- Overlap with WoS- 35,554 (50.91% of Scopus), 34,280 DOIs are non-overlapping (49.09% of Scopus)
- Overlap with OpenAlex- 38,005 (54.42% of Scopus), 31,829 are non-overlapping (45.58% of Scopus)
- Unique DOIs in Scopus- 16,527 (23.67% of Scopus)

OpenAlex- total pre-processed DOIs = 2,14,873

- Overlap with WoS- 39,718 (18.48% of OpenAlex), 1,75,155 are non-overlapping (81.52% of OpenAlex)
- Overlap with Scopus- 38,005 (17.69% of OpenAlex), 1,76,868 are non-overlapping (82.31% of OpenAlex)
- Unique DOIs in OpenAlex- 1,57,402 (73.25% of OpenAlex)

Distribution of Publications across different SDGs in the three databases

The publication records classified under one or more SDGs were analysed for the distribution across SDGs. The proportionate share of publication records classified in each SDG in all the three databases was computed (**Figure 2**). Across the three databases, SDG 03 has the highest number of records (WoS: 45,316, Scopus: 31,587 and OpenAlex: 49,951) and SDG 07 has the second most number of records (WoS: 10,539, Scopus: 13,881 and OpenAlex: 33,855). In the third position however, WoS has SDG 13 with 9,703 records, Scopus has SDG 09 with 8,326 records and Open Alex has SDG 02 with 21,830 records. SDGs 01, 04, 08, 10, 16 and 17 have less than 1000 records classified in WoS which is less than 1% of total records.

A bird's eye view picture of publication records shows differences in the proportional share of records mapped to the SDGs among the three databases **(Figure 2)**. Major variations are seen in case of WoS where more than half of the mapped records have been associated with SDG 3 (Good Health and Well Being) followed by SDG 07 (13.3%) and SDG 13 (12.2%); while SDG 4 (Quality Education), SDG 10 (Reduced Inequalities), SDG 16 (Peace, Justice and Strong Institutions) and SDG 17 (Partnerships for the Goals) have significantly lower number of associated records. In Scopus, SDG 3 has about 45.2% records mapped followed by SDG 7 (19.8%) and SDG 9 (11.9%). Remaining records are mapped mainly to SDG 2, 6, 8, 11, 12, and 13. On the other hand, in Open Alex, the highest percentage of records mapped to one SDG is 23.2% for SDG 3, followed by SDG 7 with 15.7%. This variation in the mapping of records in the databases may be an indication that in addition to their coverage variations, the schemes they utilize for SDG mapping are also different.

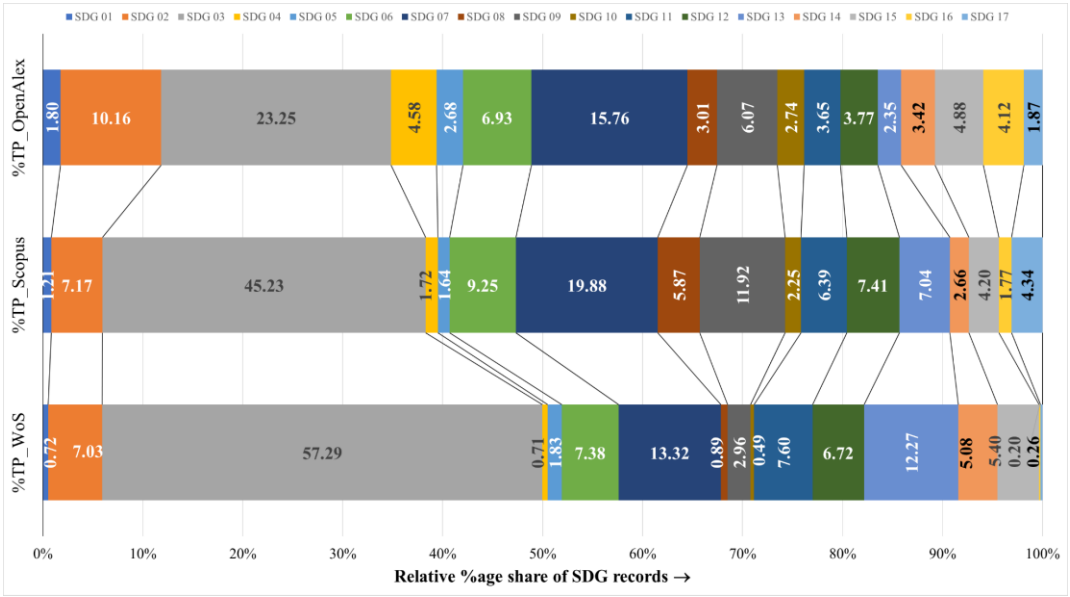
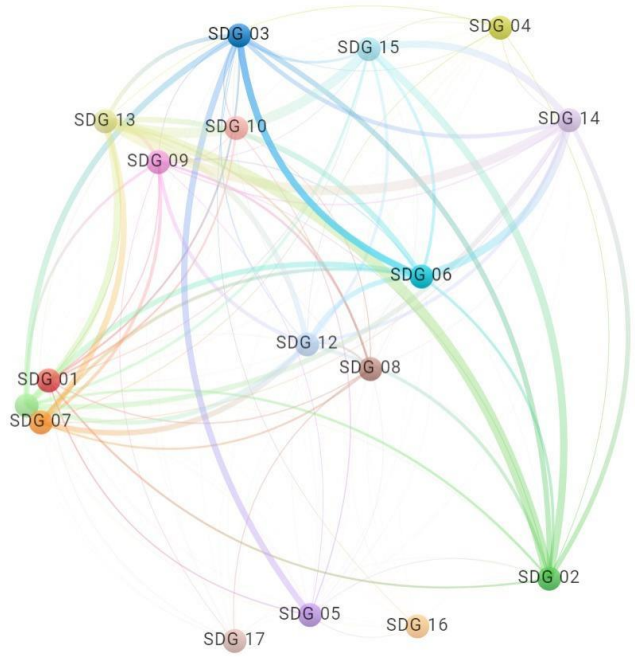


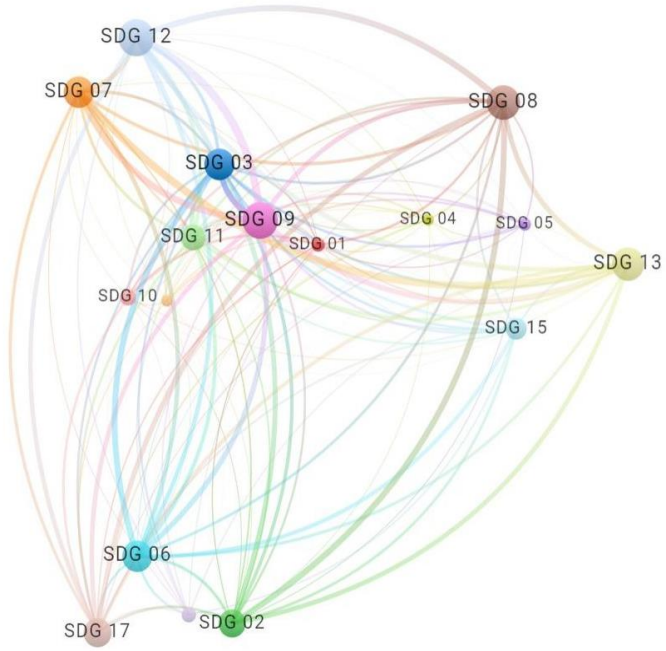
Figure 2. SDG wise tagging distribution of publications across three databases.

SDG mapping linkages across three databases

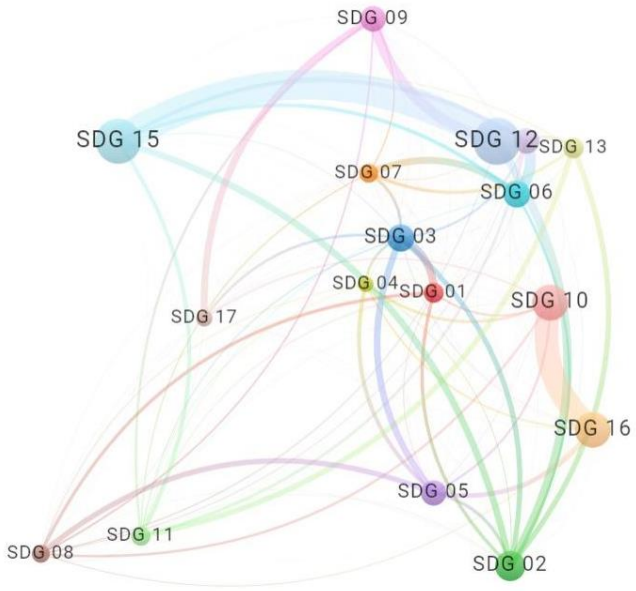
The mapping of records to each SDG in the selected database was plotted on a network map with vertices representing the total number of records assigned to the selected SDG and edges connecting two vertices showing the number of publication records which are mapped to both the SDGs **(Figure 3)**. As the edges in these maps visualise only the records mapped to multiple SDGs they provide a clearer and more detailed view into the mapping approaches of each database. In the case of WoS, edges from SDG 13 - 14 (edge weight, Wt. 2334), SDG 13 - 15 (Wt. 2542), SDG 13 - 02 (Wt. 2890), and SDG 13 - 11 (Wt. 1574), from SDG 14 - 15 (Wt. 1538) and, from SDG 15 - 02 (Wt. 1657) have the highest weights. In the case of Scopus, records are mapped together mostly between SDG 09 - 12 (Wt. 1822), SDG 09 - 07 (Wt. 1361), SDG 07 - 13 (Wt. 1663), and SDG 08 - 12 (Wt. 1293). While in Open Alex, edges between SDG 12 - 15 (Wt. 539), 09 (Wt. 122), SDG 10 - 16 (Wt. 382) and SDG 14 - 06 (Wt. 126) show a majority of records mapped together. Thus, a variation is also seen in SDG mapping linkages across the databases with common records mapped pairs among SDG 02, 11, 13, 14 and 15 in WoS; SDG 07, 08, 09, 12 and 13 in Scopus and, SDG 06, 09, 10, 12, 14, 15 in Open Alex.



(a) Web of Science



(b) Scopus



(c) Open Alex

Figure 3. SDG Interlinkages across three databases. (created using VoS Viewer)

Divergence of SDG mapping across the three databases

A further in-depth analysis of records in each database was conducted to find out records associated with multiple SDGs. This was done to analyze the focused or divergent nature of the mapping. Across the three databases, different relative proportions of records were observed to have mapping in multiple SDGs. The different pie charts plotted for each of WoS, Scopus and OpenAlex databases in **Figures 4a**, **4b** and **4c** provide further insights into SDG classification of DOIs in these databases. The pie charts depict the percentage share of DOIs classified into more than single SDG by a database.

In Open Alex, most of the records (98.92%) are mapped to only one SDG, in WOS and Scopus, 17.08% and 24.03% records are mapped to two or more SDGs. The results have been computed w.r.t. to 79,099 unique DOIs in WoS, 69,834 unique DOIs in Scopus and 214,873 unique DOIs in OpenAlex across all SDGs. From these charts, it can be seen that OpenAlex has the maximum percentage of DOIs tagged into a single SDG (98.93%) followed by WoS that classifies 82.92% of DOIs tagged into a single SDG and then Scopus which classifies approx. 76% of DOIs into a single SDG. It is observed that in WoS, the maximum limit to which the DOIs are tagged into more than single SDG is 9 while for Scopus it is 11 i.e. at most a DOI is tagged into 9 SDGs by WoS and 11 SDGs by Scopus. For OpenAlex, the maximum number of SDGs a DOI is tagged into other SDGs is 6 that too is a very minor percentage. Mostly, DOIs are classified into a single SDG by OpenAlex followed by DOIs classified into two SDGs while a very little percentage of DOIs are classified into three (3), four (4) and six (6) SDGs.

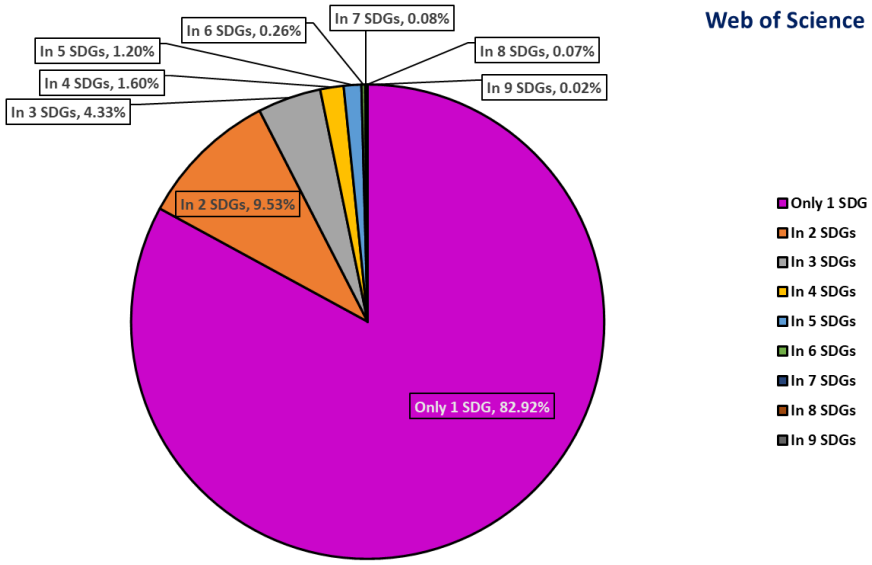


Figure 4a. Proportionate share of Publication Records mapped in single and multiple SDGs (Web of Science- 79,099 unique DOIs).

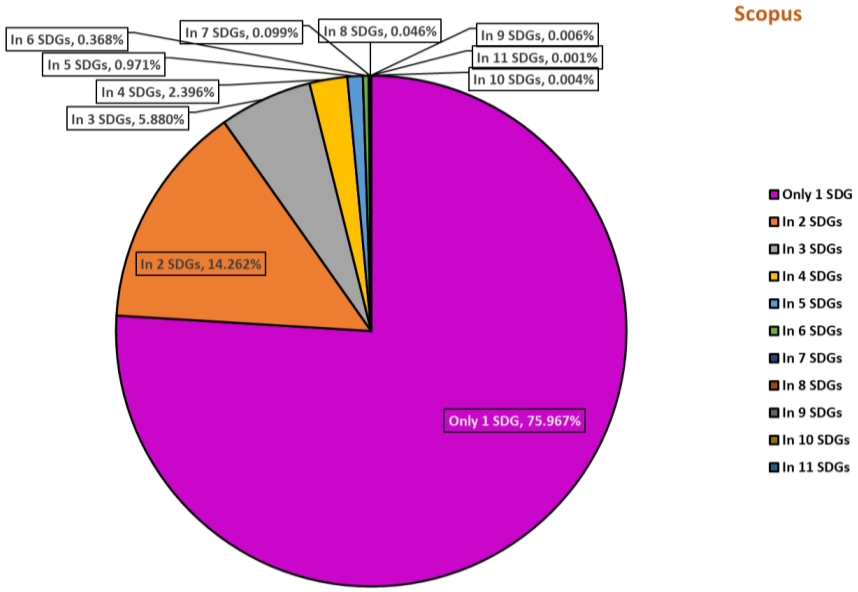


Figure 4b. Proportionate share of Publication Records mapped in single and multiple SDGs (Scopus-69,834 unique DOIs).

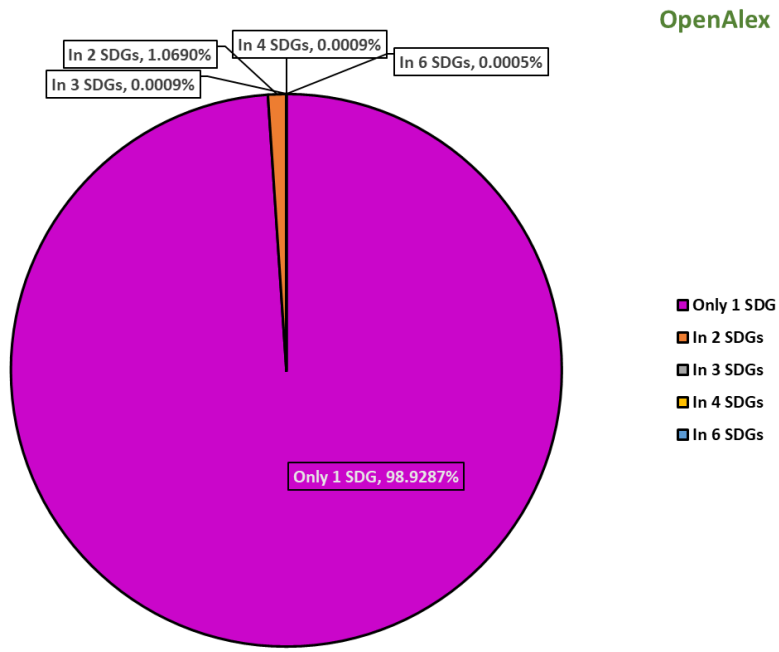


Figure 4c. Proportionate share of Publication Records mapped in single and multiple SDGs (OpenAlex- 214,873 unique DOIs).

The divergence of classification of publication records into different SDGs in each database has been analyzed next **(See Appendix)**. The methodology for computation of the divergence values for the 17 SDGs in each database is explained in the relevant section above. Results indicate that the divergence values for WoS range from 1.16 to 3.3, for Scopus it ranges from 1.23 to 3.2 while for OpenAlex it ranges from 1.00 to 1.09. Minimum value of divergence is observed for SDG 03 in both WoS and Scopus while OpenAlex has the minimum value of divergence value for SDG 07. However, the maximum divergence value is observed for SDG 14 in WoS, for SDG 01 in Scopus and for SDG 12 in OpenAlex. This indicates that a lower proportion of DOIs classified in SDG 03 in WoS and Scopus are classified into other SDGs and a lower proportion of DOIs classified in SDG 07 in OpenAlex are classified into other SDGs. Similarly, DOIs classified in SDG 14 in WoS are the ones to have been classified into more SDGs as compared to DOIs classified under other SDGs. DOIs classified in SDG 01 in Scopus are the ones to have also been classified into more SDGs as compared to DOIs classified under other SDGs. Higher proportion of DOIs classified in SDG 12 in OpenAlex have been classified into other SDGs. Also, WoS and Scopus have a mean divergence score of 2.36 and 2.38 respectively while OpenAlex has a mean divergence score of 1.03.

The range of scores and their depiction in the box plot (**Figure 5**) indicates that the SDG classification in WoS is more divergent than that in Scopus while the least divergent SDG classification is shown by OpenAlex. This implies that publication records classified under a particular SDG in OpenAlex are less likely to be mapped to other SDGs too. These differences can be attributed to the difference in the

schemes of SDG classification deployed in each database. This result further supplements the divergence of the databases in terms of SDG classification, where OpenAlex is found to be the least divergent one, followed by WoS, while Scopus is the most divergent in which more publication records are classified into multiple SDGs.

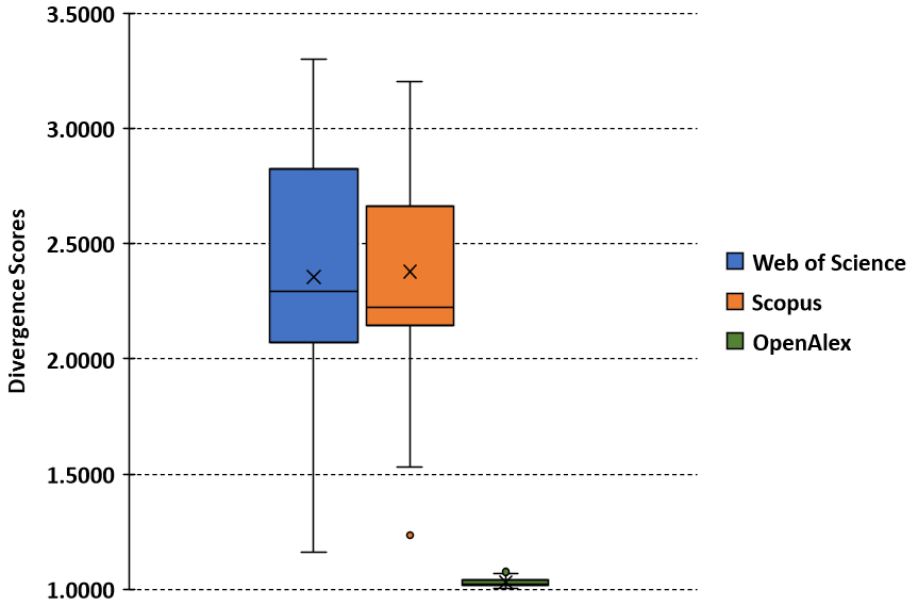


Figure 5. Range and Mean of Divergence Values for WoS, Scopus and OpenAlex.

Discussion

The study has analysed SDG mapping of publication records across the three databases, more specifically on the parameters of variations, interlinkages, and classification divergence. The results suggest variations in linking of research publication metadata to the 17 SDGs. Across the studied databases, the proportion of publications mapped with SDGs varies significantly indicating that the classification approaches used by them put varied attention to the different SDGs. *While 63% of records in Open Alex and WoS are mapped to SDGs, only 35% publication records in Scopus are associated with an SDG*. This could be a result of the keyword-based search approach used in Scopus (Bedard-Vallee *et al.*, 2023) which would be limited and slow to adapt to the continuously changing SDG research landscape. As a result, the coverage of mapped records in Scopus will only improve when the search queries are augmented. In contrast, Open Alex and WoS have the classification done at the backend level providing them flexibility to modify the classification criteria iteratively. This however places a restriction on the users of the database.

The three databases have a significant number of records which are not mapped to the same SDGs across them. An in-depth analysis of SDG mapping of records underlined the variations in database's approaches. In terms of individual SDGs, the highest number of records categorised under SDG 03 (Good Health and Wellbeing)

across the databases but the proportionate contributions range from 23% in Open Alex to 57% in WoS. *In each database, the proportion of records mapped to individual SDGs varies significantly when compared to the other databases.* This variation may be due to the differences in mapping approaches used in these databases, namely keywords-based search query in Scopus, keyword based filters in WoS, and machine learning based in Open Alex. Upon exploring the keywords used in SDG classification approach for SDG 3 it was observed that Scopus Query has 2391 words, WoS Criteria is based on 32 words, and Open Alex's ML model is trained based on results from 151 Keywords further highlighting the differences in approaches of each database. These observations conform with the earlier seen presence of variation in SDG mapping of records across the databases by some previous studies (Armitage *et al.*, 2020, Purnell, 2022, Wang *et al.*, 2023). A more detailed analysis of the exact operation of each approach would reveal a further more detailed picture.

The mapping of records across the three databases shows differences in SDG classification linkages further reinforcing the finding as above. *While approaches used by Scopus and WoS have higher affinity of characterising/mapping records in multiple SDGs, in case of Open Alex, most records are assigned to a single SDG.* WoS has divergence values ranging from 1.1 to 3.4 with about 17% of records assigned to two or more SDGs. For Scopus this range is 1.5 to 3.2, with about 24% of records mapped to two or more SDGs. However, in Open Alex the divergence values range from 1.0 to 1.3 with only 1% records having been mapped to more than one SDGs. This suggests that Open Alex is more focused and conclusive to favour individual SDGs as compared to the other two databases.

Finally, *the network maps show the interlinkages between different SDGs by looking at the individual records and their mapping with multiple SDGs.* These linkages also vary across databases indicating differences in establishing linkages between the SDGs. This is possibly a result of the fact that individual SDG classifications are drawn separately without much consideration to the commonalities between them based on their targets and application areas.

Conclusion

This study has presented a detailed analysis of the variations that exist between the SDG mapping of records in three databases, namely, WoS, Scopus and Open Alex. Variations at the level of record volume, distribution across SDGs, divergence in mapping and linkages between SDGs are explored and presented. It is observed that the approach used by Open Alex classified publication records under fewer SDGs as compared to Scopus and WoS. This is indicated by the divergence score computed using the approach proposed in this article. There are also significant differences in SDG mapping linkages across the three databases. The analysis suggests that there are not only coverage level variations across the three databases, but there are also more methodological differences in SDG mapping schemes of the databases. The results of the present study bring out more detailed insight into the SDG mapping in the three databases.

The present study, however, has some limitations as well. It uses publication data for research output for the year 2023 from one country 'India'. However, owing to a significantly large scale of analysed data in the study, it is likely that the overall findings in terms of variability in SDG mapping across databases would still be visible in other large data samples. Further, only three databases are compared, the quantitative as well as the qualitative results are indicative of variations in these databases. It may be useful and interesting to study such variations in other databases (such as Dimensions, Google Scholar, Lens.org etc.) as well. Already, some studies have proposed alternative modes of classification of publication records into different SDGs (Wulff et al., 2023). Additionally, there are no studies to evaluate the accuracy of mapping publication records to SDGs at a more microscopic level, and therefore such a study can be conducted through a user-based annotation and evaluation.

References

Armitage, C. S., Lorenz, M., & Mikki, S. (2020). Mapping scholarly publications related to the Sustainable Development Goals: Do independent bibliometric approaches get the same results?. *Quantitative Science Studies*, **1(3)**, 1092-1108.

Bautista-Puig, N., Aleixo, A. M., Leal, S., Azeiteiro, U., & Costas, R. (2021). Unveiling the research landscape of sustainable development goals and their inclusion in higher education institutions and research centers: major trends in 2000–2017. *Frontiers in Sustainability*, *2*, 620743.

Hajikhani, A., & Suominen, A. (2022). Mapping the sustainable development goals (SDGs) in science, technology and innovation: application of machine learning in SDG-oriented artefact detection. *Scientometrics*, *127*(11), 6661-6693.

IISD, Sustainable Development, International Institute for Sustainable Development (IISD), (2021). Accessed on May 27, 2023 from https://www.iisd.org/about-iisd/sustainable-development.

Kashnitsky, Y., Roberge, G., Mu, J., Kang, K., Wang, W., Vanderfeesten, M., ... & Labrosse, I. (2024). Evaluating approaches to identifying research supporting the United Nations Sustainable Development Goals. *Quantitative Science Studies*, 1-18.

Moyer, J. D., & Hedden, S. (2020). Are we on the right path to achieve the sustainable development goals?. *World Development*, *127*, 104749.

Ottaviani, M., & Stahlschmidt, S. (2024). On the performativity of SDG classifications in large bibliometric databases. *arXiv preprint arXiv:2405.03007*.

Purnell, P. J. (2022). A comparison of different methods of identifying publications related to the United Nations Sustainable Development Goals: Case study of SDG 13—Climate Action. *Quantitative Science Studies*, **3(4)**, 976-1002.

Sachs, J. D. (2012). From millennium development goals to sustainable development goals. *The lancet*, **379**(9832), 2206-2211.

Singh, A., Kanaujia, A., Singh, V. K., & Vinuesa, R. (2024). Artificial intelligence for Sustainable Development Goals: Bibliometric patterns and concept evolution trajectories. Sustainable Development, 32(1), 724-754.

Singh, A., Kanaujia, A., & Singh, V. K. (2022). Research on sustainable development goals: how has indian scientific community responded?. *Journal of Scientific & Industrial Research*, *81*(11), 1147-1161.

Sweileh, W. M. (2020). Bibliometric analysis of global scientific literature on vaccine hesitancy in peer-reviewed journals (1990–2019). *BMC public health*, *20*, 1-15.

United Nations (2020). The Sustainable Development Goals Report 2020, Accessed from https://sdgs.un.org/publications/sustainable-development-goals-report-2020-24686 on 17 April 2025.

Wang, W., Kang, W., & Mu, J. (2023). Mapping research to the sustainable development goals (sdgs). *Preprint]. https://doi. org/10.21203/rs*, *3*.

Wulff, D. U., Meier, D. S., & Mata, R. (2024). Using novel data and ensemble models to improve automated labeling of Sustainable Development Goals. *Sustainability Science*, 1-15.

Appendix

Records linked to each SDG in different databases

Table A1. Web of Science

WoS	SDG 01	SDG 02	SDG 03	SDG 04	SDG 05	SDG 06	SDG 07	SDG 08	SDG 09	SDG 10	SDG 11	SDG 12	SDG 13	SDG 14	SDG 15	SDG 16	SDG 17	Sum	TP_WoS	Divergence Score
SDG 01	568	251	267	0	158	0	0	103	134	223	0	0	73	0	63	5	14	1859	568	3.27
SDG 02	251	5557	1001	77	30	430	10	0	67	0	339	511	2890	885	1657	0	0	13705	5557	2.47
SDG 03	267	1001	45316	44	1320	1043	10	13	77	158	1253	151	719	762	364	65	0	52563	45316	1.16
SDG 04	0	77	44	564	2	0	0	0	3	11	0	80	80	77	77	0	0	1015	564	1.80
SDG 05	158	30	1320	2	1444	0	0	13	62	133	0	0	0	0	0	42	0	3204	1444	2.22
SDG 06	0	430	1043	0	0	5835	508	14	118	0	1026	940	1119	1872	481	4	0	13390	5835	2.29
SDG 07	0	10	10	0	0	508	10539	243	506	0	884	1084	1233	71	73	0	0	15161	10539	1.44
SDG 08	103	0	13	0	13	14	243	701	322	96	14	243	243	14	14	0	118	2151	701	3.07
SDG 09	134	67	77	3	62	118	506	322	2341	72	387	619	576	138	81	0	77	5580	2341	2.38
SDG 10	223	0	158	11	133	0	0	96	72	384	7	0	7	0	0	2	9	1102	385	2.86
SDG 11	0	339	1253	0	0	1026	884	14	387	7	6015	841	1574	980	430	4	0	13754	6015	2.29
SDG 12	0	511	151	80	0	940	1084	243	619	0	841	5313	1053	566	479	0	0	11880	5313	2.24
SDG 13	73	2890	719	80	0	1119	1233	243	576	7	1574	1053	9703	2334	2542	4	0	24150	9703	2.49
SDG 14	0	885	762	77	0	1872	71	14	138	0	980	566	2334	4020	1538	4	0	13261	4020	3.30
SDG 15	63	1657	364	77	0	481	73	14	81	0	430	479	2542	1538	4275	4	0	12078	4275	2.83
SDG 16	5	0	65	0	42	4	0	0	0	2	4	0	4	4	4	157	5	296	157	1.89
SDG 17	14	0	0	0	0	0	0	118	77	9	0	0	0	0	0	5	208	431	208	2.07

Table A2. Scopus

Scopus	SDG 01	SDG 02	SDG 03	SDG 04	SDG 05	SDG 06	SDG 07	SDG 08	SDG 09	SDG 10	SDG 11	SDG 12	SDG 13	SDG 14	SDG 15	SDG 16	SDG 17	Sum	TP_Scopus	Divergence Score
SDG 01	846	145	158	65	127	61	39	336	94	321	69	72	90	14	30	66	179	2712	846	3.21
SDG 02	145	5005	585	48	81	458	194	1232	317	105	173	953	828	131	420	47	404	11126	5005	2.22
SDG 03	158	585	31587	140	330	1232	473	382	674	367	871	533	383	331	377	216	328	38967	31587	1.23
SDG 04	65	48	140	1202	106	44	35	185	135	133	47	109	45	9	39	64	171	2577	1202	2.14
SDG 05	127	81	330	106	1145	22	8	206	61	275	37	21	24	7	7	266	105	2828	1145	2.47
SDG 06	61	458	1232	44	22	6459	769	408	1006	44	650	760	456	380	713	53	296	13811	6459	2.14
SDG 07	39	194	473	35	8	769	13881	537	1361	51	441	943	1663	99	228	26	501	21249	13881	1.53
SDG 08	336	1232	382	185	206	408	537	4101	880	349	347	1293	942	101	298	140	984	12721	4101	3.10
SDG 09	94	317	674	135	61	1006	1361	880	8326	138	590	1822	992	162	320	76	867	17821	8326	2.14
SDG 10	321	105	367	133	275	44	51	349	138	1572	87	85	87	18	29	128	366	4155	1572	2.64
SDG 11	69	173	871	47	37	650	441	347	590	87	4459	705	505	162	236	85	360	9824	4459	2.20
SDG 12	72	953	533	109	21	760	943	1293	1822	85	705	5172	935	198	405	84	816	14906	5172	2.88
SDG 13	90	828	383	45	24	456	1663	942	992	87	505	935	4915	194	341	49	643	13092	4915	2.66
SDG 14	14	131	331	9	7	380	99	101	162	18	162	198	194	1856	230	16	108	4016	1856	2.16
SDG 15	30	420	377	39	7	713	228	298	320	29	236	405	341	230	2933	47	255	6908	2933	2.36
SDG 16	66	47	216	64	266	53	26	140	76	128	85	84	49	16	47	1234	130	2727	1234	2.21
SDG 17	179	404	328	171	105	296	501	984	867	366	360	816	643	108	255	130	3033	9546	3033	3.15

Table A3. OpenAlex

Open Alex	SDG 01	SDG 02	SDG 03	SDG 04	SDG 05	SDG 06	SDG 07	SDG 08	SDG 09	SDG 10	SDG 11	SDG 12	SDG 13	SDG 14	SDG 15	SDG 16	SDG 17	Sum	TP_OpenAlex	Divergence Score
SDG 01	3868	24	77	0	2	1	0	23	0	8	2	0	0	0	0	0	0	4005	3867	1.04
SDG 02	24	21832	44	3	24	17	0	0	0	0	1	91	46	1	64	0	0	22147	21830	1.01
SDG 03	77	44	49993	16	63	0	18	0	0	0	4	0	0	13	2	2	17	50249	49951	1.01
SDG 04	0	3	16	9841	31	0	0	2	1	19	2	0	0	1	0	10	0	9926	9837	1.01
SDG 05	2	24	63	31	5763	0	0	47	0	14	0	0	0	1	0	39	0	5984	5763	1.04
SDG 06	1	17	0	0	0	14894	67	0	1	0	2	3	2	126	27	0	0	15140	14887	1.02
SDG 07	0	0	18	0	0	67	33864	0	7	0	6	3	22	0	0	0	0	33987	33855	1.00
SDG 08	23	0	0	2	47	0	0	6458	10	15	1	6	0	3	0	3	1	6569	6457	1.02
SDG 09	0	0	0	1	0	1	7	10	13049	0	16	122	0	0	0	0	69	13275	13049	1.02
SDG 10	8	0	0	19	14	0	0	15	0	5882	1	0	0	1	0	382	6	6328	5882	1.08
SDG 11	2	1	4	2	0	2	6	1	16	1	7849	0	45	18	30	0	0	7977	7848	1.02
SDG 12	0	91	0	0	0	3	3	6	122	0	0	8101	0	2	539	0	0	8867	8101	1.09
SDG 13	0	46	0	0	0	2	22	0	0	0	45	0	5058	23	6	0	0	5202	5058	1.03
SDG 14	0	1	13	1	1	126	0	3	0	1	18	2	23	7353	33	1	0	7576	7351	1.03
SDG 15	0	64	2	0	0	27	0	0	0	0	30	539	6	33	1049	1	0	11198	10494	1.07
SDG 16	0	0	2	10	39	0	0	3	0	382	0	0	0	1	1	8850	0	9288	8850	1.05
SDG 17	0	0	17	0	0	0	0	1	69	6	0	0	0	0	0	0	4028	4121	4028	1.02

https://doi.org/10.51408/issi2025_037

Measuring the Continuous Research Impact of a Researcher: The K_z Index

Kiran Sharma[1], Ziya Uddin[2]

[1] *kiran.sharma@bmu.edu.in*, [2]*ziya.uddin@bmu.edu.in*
School of Engineering and Technology, BML Munjal University, Gurugram, Haryana-122413 (India)

Abstract

The ongoing discussion regarding the utilization of individual research performance for academic hiring, funding allocation, and resource distribution has prompted the need for improved metrics. While traditional measures such as total publications, citations count, and the *h*-index, etc. provide a general overview of research impact, they fall short of capturing the continuous contribution of researchers over time. To address this limitation, we propose the implementation of the K_z index, which takes into account both publication impact and age. In this study, we calculated K_z scores for 376 research profiles. K_z reveals that the researchers with the same *h*-index can exhibit different K_z scores, and vice versa. Furthermore, we observed instances where researchers with lower citation counts obtained higher K_z scores, and vice versa. Interestingly, the K_z metric follows a log-normal distribution. To determine if the distribution of K_z is independent of subject discipline, we plotted the distribution for three different disciplines. Our analysis concluded that the distribution of $K_z$is indeed independent of the discipline. It highlights its potential as a valuable tool for ranking researchers and facilitating informed decision-making processes. By measuring the continuous research impact, we enable fair evaluations, enhance selection processes, and provide focused career advancement support and funding opportunities.

Introduction

Research impact (Penfield, Baker, Scoble and Wykes, 2014) is a crucial factor when evaluating the contributions of researchers (Reed, Ferre, Martin-Ortega, Blanche, Lawford-Rolfe, Dallimer and Holden, 2021). It plays a vital role in assessing the quality, significance, and reach of their work, which is instrumental in academic promotions, grant allocations, award selections, and overall career progression. Existing indices like the *h*-index and citation count are commonly used to measure research impact (Bornmann and Daniel, 2005, 2009; Egghe, 2010); however, it's important to recognize that citations may not provide a comprehensive representation of impact, especially in fields where citation practices differ or in emerging research domains with limited citation opportunities. Therefore, a more nuanced approach is necessary to capture the full extent of the research impact (de Saint-Georges and de la Potterie, 2013), considering multiple dimensions beyond traditional metrics.

Initially introduced in 2005 by Hirsch, the *h*-index is calculated based on the number of papers that have received at least *h* citations from other papers (Hirsch, 2005). The *h*-index has been subject to criticism due to its limitations in providing a comprehensive view of scientific impact (Costas and Bordons, 2007; Ding, Liu and Kandonga, 2020). It failed to capture the impact of highly cited papers (Bi, 2023).

Also, it does not take into account the number of authors in each publication (Schubert and Schubert, 2019).

However, since its introduction, the *h*-index has gained significant popularity in academia and has been commonly employed to evaluate the academic success of researchers in various areas, including hiring decisions, promotions, and grant acceptances. Despite efforts by researchers to propose alternative variants of the *h*-index (Egghe, 2006; Jin, Liang, Rousseau and Egghe, 2007; Zhang, 2009; Alonso, Cabrerizo, Herrera-Viedma and Herrera, 2010; Khurana and Sharma, 2022), the traditional *h*-index remains widely used as a performance metric in the assessment of scientists because of its simplicity. After the inception of *h*-index, many variants of *h*-index have been proposed to overcome its limitations (Alonso, Cabrerizo, Herrera-Viedma and Herrera, 2009; Batista, Campiteli and Kinouchi, 2006; Hirsch, 2019; Schreiber, 2008a, b; Todeschini and Baccini, 2016).

To overcome the limitations of *h*-index, Egghe in 2006, proposed *g*-index (Egghe, 2006) which is determined by the distribution of citations across their publications. It is determined by sorting the articles in decreasing order based on the number of citations they have received. The *g*-index is defined as the largest number g for which the top g articles collectively accumulate at least g^2 citations. This means that a researcher with a *g*-index of 10 has published at least 10 articles that collectively have received at least ($10^2 = 100$) citations. It's important to note that unlike the *h*-index, the citations contributing to the *g*-index can be generated by only a small number of articles. For example, a researcher with 10 papers, where 5 papers have no citations and the remaining five have 350, 35, 10, 2, and 2 citations respectively, would have a *g*-index of 10 but an *h*-index of 3 (as only three papers have at least three citations each).

Further, after recognizing the limitations of the *h*-index (Ding et al., 2020; Egghe, 2011), researchers have proposed various complementary measures to provide a more comprehensive assessment of research impact such as AR-index (Jin et al., 2007), *e*-index (Zhang, 2009), *p*-index (Prathap, 2010), h' -index (Zhang, 2013), h_c-index (Khurana and Sharma, 2022), etc. Van Leeuwen (2008) compared the *h*-index with various bibliometric indicators and other characteristics of researchers. Similarly, Rons and Amez (2009) proposed a new indicator named, impact validity indicator, in search of excellent scientists.

Further, the *e*-index proposed by Zhang in 2009 (Zhang, 2013), measure the excess citations received by an author's publications beyond the *h*-core. The *e*-index places a strong emphasis on highly cited papers, as it focuses on excess citations beyond the *h*-core. Jin et al. in 2007 proposed the AR-index (Jin et al., 2007) which is used to measure the citation intensity of the *h*-core (publications with at least *h* citations) while considering the age of publications. The limitation of using AR-index is that it focuses on the *h*-core without considering variations within it. Different publications within the *h*-core may have different citation counts, but the AR-index does not account for these variations. In 2010, a *p*-index introduced by Prathap

(Prathap, 2010), measure the productivity and impact by considering an author's *h*-index, total publications, and the number of citations received. The limitation of the index is that it does not consider the distribution of citations across an author's papers. It treats all papers equally and does not differentiate between highly cited and minimally cited papers.

In the study by Khurana et al. (Khurana and Sharma, 2022), an enhancement to the *h*-index is proposed to capture the impact of the highly cited paper. They introduced h_c which is based on the weight assigned to the highly cited paper. h_c has a greater impact on researchers with lower *h*-index values, particularly by highlighting the significance of their highly cited paper. However, the effect of h_c on established researchers with higher *h*-index values was found to be negligible. It is worth noting that the h_c focuses on the first highly cited paper and does not consider the impact of subsequent highly cited papers. This limitation again highlights the need for a more comprehensive measure that takes into account all the important factors contributing to research impact.

Another measure named, *L*-sequence, introduced by Liu et al. (Liu and Yang, 2014), computes the *h*-index sequence for cumulative publications while taking into account the yearly citation performance. In this approach, the *L* number is calculated based on the *h*-index concept for a specific year. Consequently, the impact of the most highly cited paper in that year may be overlooked, and papers with less than *L* citations are also not considered. Although the concept captures the yearly citation performance of all papers, it does not effectively capture the continuous impact of each individual paper. Also gathering data for the *L*-sequence can be challenging, as it requires delving into the citation history of each paper for every year.

Quantifying research impact is a multifaceted endeavour (Batista et al., 2006). There is no universally accepted metric till now to measure the continuous research impact of a researcher. Different stakeholders may prioritize different indicators, such as the number of publications, total citations, patents, etc. Measuring the continuous research impact of a researcher is crucial for granular assessment, differentiation among researchers, funding decisions, identification of emerging talent, etc. Determining an inclusive and comprehensive approach that captures the diverse dimensions of research impact remains a challenge.

Research objective

The primary objective of this study is to introduce a robust and reliable metric that can effectively capture the continuous research impact of a researcher. The aim of the proposed metric is to differentiate between two researchers who possess identical research parameters, for example; the number of publications or total citations or *h*-index, etc. In order to accomplish the stated objective, a newly introduced measure called the K_z-index has been proposed.

K_z-index

Based on the limitations of the h-index, especially *h* ignores the highly cited papers, the index h_c was proposed (Khurana et al., 2022). In index h_c a weight of the highest cited paper of an individual was computed. Following this study, the proposed K_z index serves as a tool to measure the continuous research impact of a researcher. It aims to capture the continuous and evolving contributions made by the researcher over time, considering factors such as total publications, citation count, *h*-index, and publication age.

Definition of K_z-index

K_z takes into account two important factors of research: paper impact and paper age.

1. **Impact** (k): The impact of a paper is calculated by considering two factors: the number of citations (C) it has received and its author's *h*-index.

The impact of the paper is calculated by using the following equation;

$$C \leq (h+1)^k \quad \ldots(1)$$

where $k \in R^+$ (positive real number).

2. **Age**(Δt): Δt represents the publication age in relation to the current year and can be calculated as

$$\Delta t = C_y - P_y \quad \ldots(2)$$

where C_y represents the current year and P_y represents the publication year.

Now, from equations (1) the value of "k" can be calculated and using equation (2), K_z can be calculated for every researcher as

$$K_z = \sum_{i=1}^{N} k_i' \quad \ldots(3)$$

where $k' = \frac{k}{\Delta t}$, and N is number of publications.

Equation (3) highlights the significance of K_z metric by incorporating essential research indicators, including the number of publications, total citations, year of publication, publication age, and *h*-index. This comprehensive approach ensures that all significant indicators of a researcher's work are considered, resulting in a more robust and holistic assessment of their research impact.

Advantages of K_z

Measuring the continuous research impact of a researcher is crucial for several reasons:

1. *Granular assessment:* Traditional matrices such as the number of publications, total citations, *h*-index, etc. present an overall research impact and do not have the capability to capture the ongoing progress and advancement of their work, whereas K_z can acquire a more nuanced and thorough comprehension of a researcher's contributions as they evolve over time.
2. *Differentiation among researchers:* Even if two researchers having the same *h*-index, the patterns of their research impact over time may vary significantly. Analysing their continuous research impact can uncover disparities in productivity and can provide a more comprehensive understanding of their

individual profiles. Hence, K_z allows for a more nuanced differentiation among researchers.

3. *Evaluation of long-term impact:* Researchers may experience fluctuations in their productivity and impact over their careers. Measuring continuous research impact enables the evaluation of long-term contributions. K_z has the capability of highlighting researchers who consistently generate influential work and have a lasting impact on their field.
4. *Career progression and funding decisions:* Many academic institutions, funding agencies, and hiring committees rely on research performance metrics to make decisions. K_z can provide more informed evaluations of researchers, enabling fairer assessments and enhancing the recognition of sustained excellence.
5. *Identification of emerging talent:* Continuous research impact measurement can help identify early-career researchers with promising trajectories. By recognizing their continuous growth and impact, further opportunities can be provided to nurture their potential.

Case studies of K_z

We conducted four case studies to explore the significance of K_z. Each case study involved two researchers, namely *R*1 and *R*2. The number of publications was kept constant across all cases, while the focus was on comparing the *h*-index and total citations (*TC*) of two researchers.

1. **Case I - Identical *h*-index and total citations:** Table 1 represents the first case study where we assumed that both researchers *R*1 and *R*2 have the same *h*-index and total citations count. However, despite sharing these characteristics, researcher *R*2 obtained a higher K_z score than *R*1. This difference in K_z scores can be attributed to the impact of the publication year, which played a dominant role in determining the continuous research impact of each researcher. It highlights the significance of considering the temporal aspect of research contributions when assessing the research impact on individuals.
2. **Case II - Identical *h*-index and different total citations:** In this case (Table 2), both researchers *R*1 and *R*2 have an equal number of publications and *h*-index, but they differ in their total citations count. Researcher *R*1 has one highly cited paper, while researcher *R*2 has multiple highly cited papers. Despite *R*1 having a higher total number of citations compared to *R*2, *R*2 obtains a higher K_z score. This indicates that the impact of having multiple highly cited papers outweighs the effect of a single highly cited paper in determining the continuous research impact.
3. **Case III(a) - Different *h*-index and total citations:** In this case (Table 3), both researchers have an equal number of publications but differ in their *h*-index, number of high impact papers, and total citations. Researcher *R*1 has a higher *h*-index but lower total citation count compared to *R*2. However, despite *R*1 having a lower total citation count, they obtain the highest K_z score. This highlights the importance of considering the continuous research impact captured by K_z, which takes into account not only the number of citations but also the publication age and impact of publications.

4. **Case III(b) - Different *h*-index and total citations:** In this case (Table 4), we again considered two researchers with an equal number of publications but different *h*-index, high impact papers, and total citations. Researchers *R*1 had a higher *h*-index and total citation count compared to researcher *R*2. Surprisingly, despite these differences, it was researcher *R*2 who obtained the highest K_z score. This finding suggests that the K_z score takes into account factors beyond just *h*-index and total citations, emphasizing the importance of considering the continuous impact and temporal aspects of research contributions.

Table 1. Two researchers with identical h-index and total citations.

Case I	**Researcher *R*1, *h*=4**					**Researcher *R*2, *h*=4**				
S. No	P_y	**C**	***k***	**dt**	***k′***	P_y	**C**	***K***	**dt**	***k′***
1	2014	40	2.292	9	0.255	2014	2	0.43	9	0.048
2	2015	30	2.113	8	0.264	2015	3	0.682	8	0.085
3	2016	0	0	7	0	2016	3	0.682	7	0.098
4	2017	3	0.682	6	0.114	2016	40	2.292	7	0.327
5	2018	24	1.974	5	0.395	2017	1	0	6	0
6	2019	1	0	4	0	2018	30	2.113	5	0.423
7	2020	1	0	3	0	2019	22	1.92	4	0.48
8	2021	1	0	2	0	2020	0	0	3	0
9	2022	0	0	1	0	2021	1	0	2	0
10	2022	10	1.43	1	1.431	2022	8	1.292	1	1.292
	TC = 110, K_z = 2.459					*TC* = 110, K_z = 2.753				

Table 2. Two researchers with identical h-index and different total citations.

Case II	**Researcher *R*1, *h*=4**					**Researcher *R*2, *h*=4**				
S. No	P_y	**C**	***k***	**dt**	***k′***	P_y	**C**	***K***	**dt**	***k′***
1	2014	1000	4.292	9	0.477	2014	500	3.861	9	0.429
2	2015	4	0.861	8	0.108	2015	300	3.54	8	0.443
3	2016	0	0	7	0	2016	100	2.861	7	0.409
4	2017	4	0.861	6	0.144	2016	0	0	7	0
5	2018	5	1	5	0.2	2017	2	0.43	6	0.072
6	2019	1	0	4	0	2018	50	2.43	5	0.486
7	2020	1	0	3	0	2019	1	0	4	0
8	2021	1	0	2	0	2020	3	0.682	3	0.228
9	2022	0	0	1	0	2021	1	0	2	0
10	2022	0	0	1	0	2022	0	0	1	0
	TC = 1016, K_z = 0.929					TC = 957, K_z = 2.067				

Table 3. Two researchers with different h-index and total citations where R1 has higher h-index and lower total citations than R2.

Case III (a)	Researcher *R1*, *h*=5					Researcher *R2*, *h*=3				
S. No	P_y	**C**	***K***	**dt**	***k′***	P_y	**C**	***k***	**dt**	***k′***
1	2014	90	2.511	9	0.279	2014	250	3.982	9	0.443
2	2015	80	2.445	8	0.306	2015	2	0.5	8	0.063
3	2016	20	1.672	7	0.239	2016	2	0.5	7	0.071
4	2017	3	0.613	6	0.102	2016	82	3.178	7	0.454
5	2018	24	1.773	5	0.355	2017	2	0.5	6	0.083
6	2019	2	0.386	4	0.097	2018	110	3.39	5	0.678
7	2020	3	0.613	3	0.204	2019	1	0	4	0
8	2021	3	0.613	2	0.307	2020	2	0.5	3	0.167
9	2022	2	0.386	1	0.387	2021	2	0.5	2	0.25
10	2022	23	1.75	1	1.75	2022	0	0	1	0
	TC = 250, K_z = 4.026					TC = 453, K_z = 2.209				

Table 4. Two researchers with different h-index and total citations where R1 has higher h-index and total citations than R2.

Case III(b)	Researcher *R1*, *h*=6					Researcher *R2*, *h*=4				
S. No	P_y	**C**	***k***	**dt**	***k′***	P_y	**C**	***k***	**Dt**	***k′***
1	2014	200	2.722	9	0.303	2014	2	0.43	9	0.048
2	2015	150	2.575	8	0.322	2015	2	0.43	8	0.054
3	2016	5	0.827	7	0.118	2016	3	0.682	7	0.098
4	2017	10	1.183	6	0.197	2016	1	0	7	0
5	2018	35	1.827	5	0.365	2017	280	3.501	6	0.584
6	2019	1	0	4	0	2018	2	0.43	5	0.086
7	2020	33	1.796	3	0.599	2019	40	2.292	4	0.573
8	2021	1	0	2	0	2020	70	2.639	3	0.88
9	2022	2	0.356	1	0.356	2021	2	0.43	2	0.215
10	2022	32	1.781	1	1.781	2022	50	2.43	1	2.431
	TC = 469, K_z = 4.041					TC = 452, K_z = 4.969				

Empirical study

To calculate the continuous research impact (K_z) of researchers, the research profiles of 376 individuals affiliated with Monash University, Australia were obtained. Monash University is a public research institution located in Australia, and information about the researchers can be found on their webpage at https://research.monash.edu/ en/persons/. The webpage provides the researcher's research ID and Orcid ID, which facilitated the extraction of their publication details and citations from the Web of Science database. From a pool of 6316 researchers' profiles, we selected 376 profiles across different disciplines, ensuring a range of *h*-index values $(1 \leq h \leq 112)$. The choice of databases was made based on data

availability. For each researcher ID, information regarding the publication year and the corresponding citations received were extracted. For each researcher, the *h*-index, and K_z were computed. Additionally, the overall research age or career length of the researcher was determined by subtracting the year of his/her first publication from the current year.

Comparative analysis of K_z

By using equation (3), we calculated the K_z score of 376 researchers. In Figure 1, a scatter plot depicting the relationship between K_z and career length. Each dot on the plot represents an individual researcher. The horizontal dashed line represents the median of the axis, while vertical dashed lines are used to divide the plot into three zones based on the length of the researchers' careers: early career (≤ 10 years), mid-career ($10 <$ years ≤ 20) and advanced career (> 20 years). This visualization clearly differentiates between the star performer and average performer at different career stages.

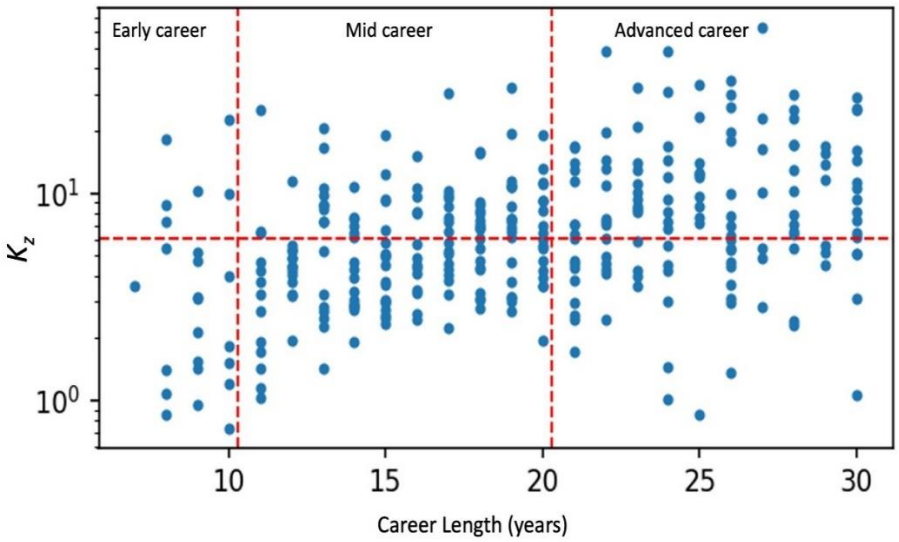


Figure 1. Scattered plot of K_z versus career length. Each dot corresponds to a researcher. The horizontal dashed line represents the median of the axis and vertical dashed lines divides the plots in three zones based on the researcher's career length.

Table 5 elucidates the significance of utilizing K_z over the *h*-index when researchers share the same *h*-index. The table presents details on career length (in years), the number of papers, total citations, *h*-index, and K_z score for a group of researchers who share a common *h*-index. Specifically, the table includes information for two sets of researchers: one set of researchers with *h*-index 25, labelled as *R*1-*R*8, and the other set of researchers with *h*-index 30 labelled as *R*9-*R*16. As highlighted earlier, the K_z score provides valuable differentiation between two researchers with the same *h*-index based on their continuous research impact. For instance, within Table 5, researchers *R*4 and *R*6 share an *h*-index of 25 and an identical number of papers (59) with total citations 2982 and 1530 respectively. However, crucially, they do not share

the same K_z score. Notably, despite *R*4 having a higher total citation count than *R*6, the former exhibits a lower K_z score.

Likewise, in the case of researcher *R*13 and *R*16, both share an *h*-index of 30. While *R*13 boasts a longer career length, a greater number of publications, and higher total citations compared to *R*16, it's noteworthy that *R*16 attains a higher K_z score. This scenario is just one among several instances of researchers depicted in the provided table. The presence of an identical *h*-index underscores its limitation in distinguishing the top-performing researcher from their peers, while K_z serves as a significant discriminator for identifying impactful researchers. This distinction emphasizes the varying impact levels among researchers. Similarly, in Table 6, profiles of researchers with the same career age are presented, yet their K_z scores differ. Consider researchers *R*9 and *R*11, who share the same career length. However, *R*9 has fewer publications and a higher number of citations and *h*-index compared to *R*11, resulting in a higher K_z score for *R*11.

Table 5. Comparative analysis among researchers having identical h-index.

S.No	Researcher ID	Career Length (Yrs)	#Papers	Total Citations	*h*-Index	K_z
R1	B-6419-2008	17	44	2415	25	5.24
R2	H-6054-2014	19	38	3433	25	6.76
R3	D-5776-2019	26	68	1984	25	6.828
R4	J-1532-2014	18	59	2982	25	7.896
R5	N-8153-2014	20	78	4217	25	9.156
R6	E-6623-2015	14	59	1530	25	10.618
R7	A-3854-2010	21	86	2034	25	11.224
R8	K-5277-2012	24	73	3783	25	11.912
R9	B-8486-2008	29	79	2851	30	4.487
R10	G-1412-2012	34	69	2816	30	5.517
R11	H-3196-2013	13	94	2538	30	8.684
R12	F-2273-2010	16	102	2627	30	10.446
R13	I-1956-2014	23	123	3797	30	11.05
R14	I-1738-2013	19	105	3306	30	11.309
R15	D-4239-2011	25	133	3343	30	12.475
R16	H-4935-2013	15	100	2945	30	18.97

Table 6. Comparative analysis among researchers having identical career length.

S.No	Researcher ID	Career Length (Yrs)	#Papers	Total Citations	*h*-Index	K_z
R1	K-5514-2018	10	9	32	4	1.043

R2	P-7354-2019	10	8	171	6	1.69
R3	I-9365-2017	10	20	287	10	3.823
R4	G-3877-2013	10	75	1189	18	9.813
R5	L-4481-2018	10	90	6012	28	22.385
R6	N-4364-2019	20	23	757	14	1.905
R7	A-4190-2009	20	32	832	14	3.795
R8	B-7556-2008	20	60	7144	27	6.847
R9	C-9764-2013	20	122	5917	42	10.995
R10	I-1587-2014	20	107	1127	18	12.88
R11	C-4319-2011	20	170	5080	39	19.088
R12	H-9193-2014	30	26	181	8	2.939
R13	P-8366-2016	30	98	5701	40	6.378
R14	B-9553-2008	30	91	6784	45	10.524
R15	H-5706-2014	30	171	4559	35	15.996
R16	A-5452-2008	30	283	26495	89	25.657
R17	I-6251-2012	30	280	58171	68	29.05

Furthermore, upon scrutinizing researchers *R*5 and *R*14 who have distinct career lengths, a noticeable disparity comes to light. Despite *R*5 being a younger researcher with a lower *h*-index than the more experienced *R*14, their research impact is effectively captured by K_z. Notably, *R*5 possesses a higher K_z score compared to *R*14.Therefore, K_z distinctly identifies impactful researchers, particularly in scenarios where researchers exhibit nearly identical numbers of publications, citations, and *h*-index.

In Table 7, we explored 11 comparative scenarios involving researchers with identical *h*-index and career length. One notable case is *S*1, where two researchers share an 8-year career length and an *h*-index of 12. Despite the similarities, the researcher with a higher total of publications and citations attains a superior K_z score. Conversely, in *S*3, where two researchers have a 13-year career and an *h*-index of 19, the one with fewer total publications but a higher citation count than the counterpart secures a higher K_z score. On the contrary, in *S*7, with a career length of 17 years and an *h*-index of 13 for both researchers, the one with more total publications but fewer citations than the other earns a higher K_z score.

This highlights that the K_z metric comprehensively considers relevant research indicators, including total publications, citation count, *h*-index, and publication age, to capture an individual's continuous impact. It's important to note that a higher K_z score cannot be solely attributed to either more total publications or higher citations count. Furthermore, one cannot conclusively assert that an individual with a higher *h*-index will always possess a higher K_z score. The K_z metric adopts a holistic approach, simultaneously considering multiple factors in the assessment of research impact.

Table 7. Comparative analysis among researchers having identical research career length (yrs) and h-index.

S.No	Researcher ID	Career Length (Yrs)	#Papers	Total Citations	*h*-Index	K_z
S1	F-9424-2013	8	37	1595	12	7.041
	O-7942-2018	8	34	454	12	5.291
S2	AAE-7279-2019	12	47	1529	15	11.122
	I-9929-2012	12	37	1236	15	4.321
S3	L-4989-2018	13	84	1875	19	20.182
	M-7607-2014	13	106	1130	19	8.26
S4	E-6431-2011	14	16	508	8	4.057
	N-1676-2017	14	14	726	8	2.771
S5	A-7222-2013	14	28	608	14	6.299
	L-1320-2019	14	23	875	14	3.264
S6	K-7419-2014	15	52	482	11	2.845
	G-1470-2011	15	36	351	11	4.741
S7	O-9174-2014	17	36	708	13	4.444
	J-5651-2016	17	16	857	13	2.173
S8	Q-9068-2018	18	47	2034	21	7.279
	H-4554-2014	18	53	1462	21	8.99
S9	F-6776-2014	18	159	1843	23	15.62
	H-8387-2012	18	78	1798	23	8.635
S10	F-4112-2014	22	18	617	13	2.402
	C-6296-2014	22	35	1456	13	4.842
S11	C-2440-2013	28	38	6087	27	2.401
	N-5018-2017	28	87	2588	27	7.02

Probability distribution of K_z

Figure 2 presents a graphical representation of the plot for $\log(K_z)$, which exhibits a mean value of μ and a standard deviation of σ. This plot is compared to the normal distribution with the same mean and standard deviation. The overlapping nature of the two plots suggests that the variable K_z follows a log-normal distribution.

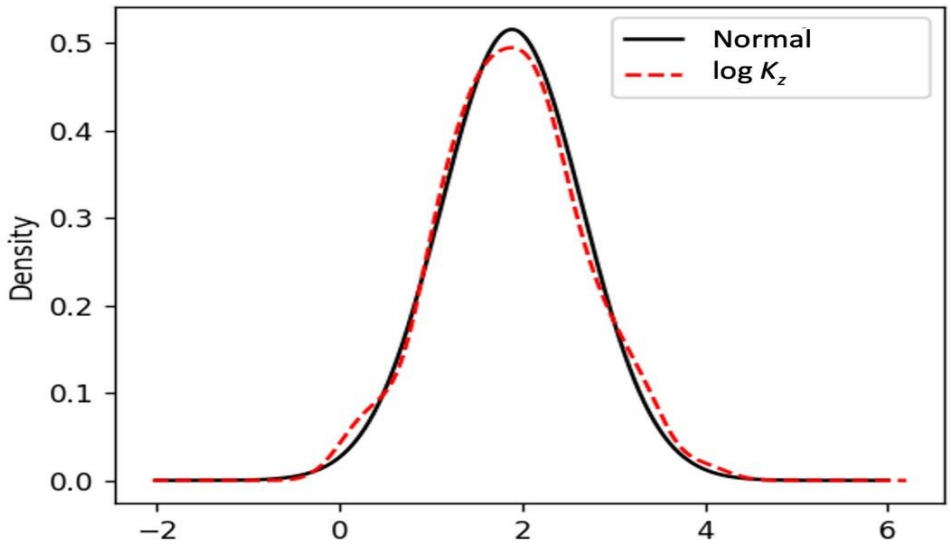


Figure 2. Distribution of $\log(K_z)$ (dashed) versus normal distribution (solid) with same μ and σ.

To confirm this observation, a "Goodness of Fit" test was conducted using the χ^2 distribution. The objective of the Goodness of Fit Test was to assess the suitability of the null hypothesis that states "the distribution of $\log(K_z)$ conforms well to a normal distribution." The test was executed in the following manner:
The logarithm of the values of K_z was computed, and these values were then classified into seven distinct classes, taking into account the mean ($\mu = 0.78787$) and standard deviation ($\sigma = 0.37448$). Subsequently, the observed frequencies (O_i) for each class were determined. To obtain the expected frequencies (E_i), the entire dataset consisting of 376 observations was subjected to calculations based on the normal distribution. The specific calculations and their results are provided in Table 8.

Table 8. Goodness of fit test.

Classes	**Observed Frequencies (O_i)**	**Expected frequencies (E_i) for $\mathcal{N}(\mu, \sigma)$**
$\log(K_z) < \mu - 1.5\sigma$	14	25
$\mu - 1.5\sigma \leq \log(K_z) < \mu - \sigma$	34	35
$\mu - \sigma \leq \log(K_z) < \mu - 0.5\sigma$	57	56
$\mu - 0.5\sigma \leq \log(K_z) < \mu + 0.5\sigma$	157	144
$\mu + 0.5\sigma \leq \log(K_z) < \mu + \sigma$	57	56
$\mu + \sigma \leq \log(K_z) < \mu + 1.5\sigma$	29	35
$\log(K_z) \geq \mu - 1.5\sigma$	28	25
Total	376	

The χ^2 value was computed using the formula $\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ and yielded a value of 7.466. As the calculated χ^2 value is smaller than the critical value $\chi^2_{(6,0.05)} = 12.592$, we cannot reject the null hypothesis at a significance level of 0.05. Therefore, we can conclude that $\log(K_z)$ is a suitable fit for the normal distribution.

Analysis of K_z distribution across Physical Science, Agriculture, and Engineering & Technology Domains

To compare the Kz score across different disciplines, authors profiles of 931 researchers from Physical Science, 432 from Agriculture, and 887 from Engineering & Technology domains has been analyzed. The Scopus ID of all authors has been extracted from Indian Scholars profile database, named VIDWAN (https://vidwan.inflibnet.ac.in/). Then the complete profile of authors with their publication details and citations has been extracted from Scopus database. Further, the corresponding $\boldsymbol{K_z}$ values are computed using equation (3). Hence the domain wise distribution of $\log(K_z)$ is plotted and shown in the Figure 3. Form the figure it is observed that $\boldsymbol{K_z}$ follows the log-normal distribution in all the three research domains.

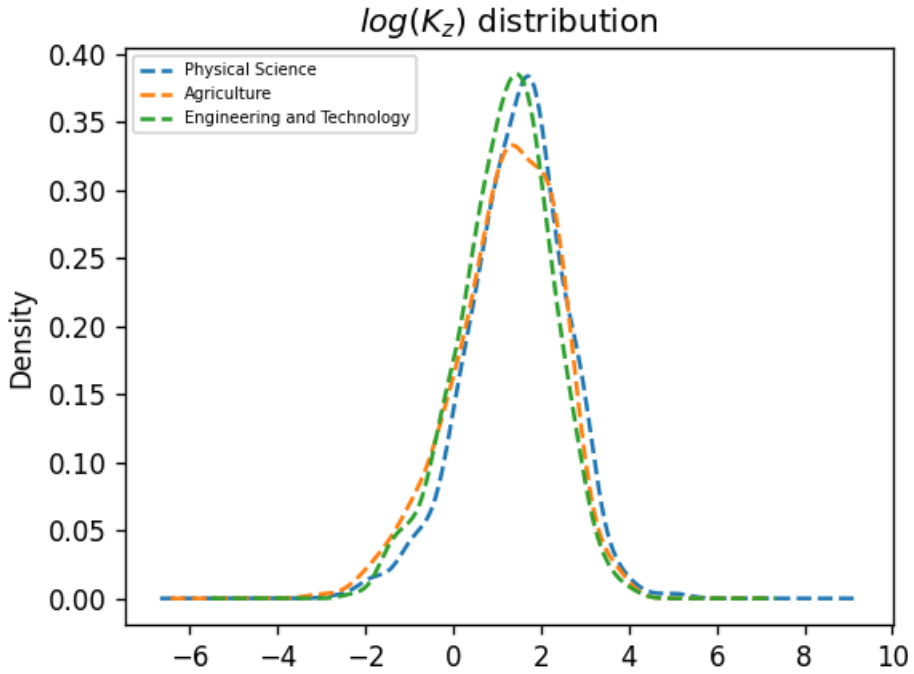


Figure 3. Distribution of $\log(K_z)$ for Physical Science, Agriculture, and Engineering & Technology domains.

It was also observed that the variance of $\log(K_z)$) is consistent across these disciplines. To statistically confirm this, Bartlett's test is applied as follows:

Null Hypothesis: *All variances are equal*

Alternative Hypothesis: *At least one variance is different*

Significance Level: *0.05*

The test resulted in a χ^2 value of 5.77 and a p-value of 0.056, supporting the equality of variances. With equal variances confirmed, ANOVA was used to determine if the mean of $\log(K_z)$ differs among the disciplines. The ANOVA results are shown in Table 9. The p-value being less than the significance level (0.05) indicates that the mean $\log(K_z)$ significantly differs across the three disciplines.

Table 9. Data Summary and ANOVA.

DATA SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Agriculture	432	224.6054	0.51992	0.258682
Engineering	887	453.1929	0.510928	0.212634
Physical Science	931	577.7204	0.620537	0.23162

ANOVA

Source of Variation	*SS*	*Df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	6.232933	2	3.116466	13.58977	1.36E-06	2.99973
Within Groups	515.2919	2247	0.229324			
Total	521.5248	2249				

To further investigate if any two domains have similar mean of $\log(K_z)$ values, Tukey's Honest Significant Difference (HSD) and Fisher's LSD tests were applied. The summaries of these tests are provided in Tables 10-13. Both tests reveal that the mean $\log(K_z)$ is the same for Agriculture and Engineering & Technology domains, but differs for the Physical Science domain.

Tukey Pairwise Comparisons

Table 10. Grouping Information Using the Tukey Method and 95% Confidence.

Factor	**N**	**Mean**	**Grouping**	**Remark**
Physical Science	931	0.6205	A	*Means that do*
Agriculture	432	0.5199	B	*not share a letter*
Engineering & Technology	887	0.5109	B	*are significantly different*

Table 11. Tukey Simultaneous Tests for Differences of Means.

Difference of Levels	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
Engineering - Agriculture	-0.0090	0.0281	-0.32	0.945
Physical Sci - Agriculture	0.1006	0.0279	3.61	0.001
Physical Sci - Engineering	0.1096	0.0225	4.88	0.000

Fisher Pairwise Comparisons

Table 12. Grouping Information Using the Fisher LSD Method and 95% Confidence.

Factor	N	Mean	Grouping	Remark
Physical Science	931	0.6205	A	*Means that do not share a letter are significantly different*
Agriculture	432	0.5199	B	
Engineering & Technology	887	0.5109	B	

Table 13. Fisher Individual Tests for Differences of Means.

Difference of Levels	Difference of Means	SE of Difference	T-Value	Adjusted P-Value
Engineering - Agriculture	-0.0090	0.0281	-0.32	0.749
Physical Sci - Agriculture	0.1006	0.0279	3.61	0.000
Physical Sci - Engineering	0.1096	0.0225	4.88	0.000

Identification of top contributors and low contributors

In the case of a normal distribution, the middle 50% of the data is encompassed within a range of +0.67 and -0.67 standard scores from the mean. Consequently, researchers in the top 25% satisfy the condition $K_z \geq e^{(\mu-0.67\sigma)}$, while researchers in the bottom 25% satisfy the condition $K_z \leq e^{(\mu-0.67\sigma)}$. Similarly, using the properties of normal distribution, the α% of top and bottom performers can be identified. Unlike previous indices such as the h, g, e, h_c, etc., the K_z-index allows for the identification of both top and bottom contributors. This categorization based on K_z scores can be beneficial for universities, scientific communities, and research funding agencies in identifying significant contributors.

Discussion and conclusion

In this study, we have discussed various research indicators, including total publications, citations count, *h*-index, etc., commonly used to measure the impact of research. While total publications, citation count, and *h*-index are commonly used indicators to assess research impact, they have some limitations when considered individually.

Some of the limitations when considering the research indicators alone are highlighted below:

1. *Total publications*: Relying solely on the number of publications can be misleading, as it does not consider the quality or impact of those publications. Quantity alone does not reflect the significance or influence of a researcher's work.
2. *Citations count*: While citation count is a useful indicator of the influence and visibility of a researcher's work, it can be influenced by factors such as the field of study, publication age, and citation practices within the research community. Additionally, self-citations can artificially inflate citation counts and impact assessments.
3. *h-index:* The *h*-index takes into account both the number of publications and their corresponding citations. However, it does not differentiate between highly cited publications and those with fewer citations. A researcher with a few highly influential papers can have the same *h*-index as someone with many moderately cited papers. Additionally, *h*-index ignores all the papers which are cited less than *h*.
4. *Temporal considerations:* Individual metrics may not capture the continuous progress and development of a researcher's work over time. They provide a snapshot of impact at a specific moment and may not reflect the long-term contributions or evolving research trajectory.

To overcome these limitations and capture the dynamic nature of research impact, it is essential to consider multiple indicators and employ comprehensive assessment approach. We attempted to address above mentioned issues and proposed an index named K_z index, which incorporates various factors to provide a more nuanced understanding of research impact. This study focuses on certain drawbacks of the *h*-index, particularly its exclusion of papers with citations below the *h*-index value and those exceeding it. To illustrate, if the *h*-index is 10, papers with citations below 10 are deemed to have no impact on the author's contribution and are consequently excluded from the *h*-index calculation. Moreover, whether a paper has 20, 30, or 100 citations, they all contribute equally to the *h*-index value, which remains fixed at 10. In contrast, our proposed index, K_z, considers all papers regardless of citations being higher or lower than the author's *h*-index. The paper explicitly delineates scenarios where a high *h*-index alone may not necessarily indicate an active researcher. Additionally, we take into account the time of publication and the popularity of papers over both long and short periods to gauge the author's contribution to the research community. The distribution of K_z is field independent as well as takes into

account the temporal aspect of the work. Unlike other research indicators, K_z takes into account not only the total publications and citations count but the age of the publications too. Our results demonstrate how K_z can effectively differentiate between two potential researchers who may have the same *h*-index, citations count, or career length. By incorporating K_z into the evaluation process, we can better assess the research dynamics of an individual and gain insights into their continuous impact over time.

To conclude, K_z holds the potential to serve as a superior measure for capturing the impact of individuals, institutions, or journals. Its comprehensive consideration of various research indicators (total citations, total publications, *h*-index, etc.) allows a more nuanced assessment of research impact. Further K_z can be utilized as a ranking method to evaluate and rank researchers within an institution based on their research impact. Similarly, institutions and journals can be compared and ranked according to their research impact. This information can be valuable in decision-making processes, as funding agencies, research award committees and hiring bodies can leverage the power of K_z to rank potential candidates within a specific field. It provides a standardized tool to assess and compare the impact of research entities, facilitating more informed decisions and promoting recognition based on research excellence.

There are some challenges and limitations associated while computing the K_z metric too.

1. *Data availability and accuracy:* Different databases may have variations in the coverage of publications and citations, potentially leading to incomplete or inconsistent data. Obtaining accurate and comprehensive data from various sources can be a challenge.
2. *Data quality and reliability:* The accuracy and reliability of the data gathered from different data sources, used for computing K_z are crucial as inaccurate or incomplete data can result in misleading assessments of research impact.
3. *Self-citation manipulation:* The issue of self-citation manipulation, where researchers excessively cite their own work to inflate their impact metrics, can pose a challenge as detecting such manipulations requires careful scrutiny and data filtering techniques.
4. *Special case for citation 0 or 1:* As mentioned earlier, K_z works fine for all the cases of papers with more than 1 citation. As K_z is computing the continuous research impact of an author, therefore the papers with zero and one citation have been considered as having no impact. The proposed index K_z is the summation of all the individual *k* values divided by the time interval, therefore the papers with zero citation or 1 citation do not seem to have much significance in the continuous research impact of a particular author. In previously published studies by Khurana et al. (Khurana and Sharma, 2022), the authors have shown such cases as limiting cases.
5. *Fractional citations:* For multiauthor publication, the proposed index does not provide the fractional weightage to citations (Bi, 2023). At present, each

individual in the multiauthor publication received the full citation while computing the K_z-score.

As discussed, it can be inferred that the K_z index is a comprehensive mathematical function that considers multiple factors to assess the impact of a researcher. These factors include the researcher's total publications, the citation count of each paper, the researcher's *h*-index, and the age of publication. The K_z index recognizes influential papers which often receive citations at a faster rate, indicating a greater impact, and therefore assigns them higher weight in impact evaluation. By considering these aspects, the K_z index tends to yield higher values in cases where a researcher has made significant contributions that have garnered substantial citations.

Data Availability Statement

The datasets generated during and/or analysed during the current study along with python codes are available from the corresponding author on reasonable request.

Conflict of interest

The author declares no conflict of interest.

References

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F., 2009. h-index: A review focused in its variants, computation and standardization for different scientific fields. Journal of informetrics 3, 273–289.

Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F., 2010. hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices. Scientometrics 82, 391–400.

Batista, P.D., Campiteli, M.G., Kinouchi, O., 2006. Is it possible to compare researchers with different scientific interests? Scientometrics 68, 179–189.

Bi, H.H., 2023. Four problems of the h-index for assessing the research productivity and impact of individual authors. Scientometrics 128, 2677– 2691.

Bornmann, L., Daniel, H.D., 2005. Does the h-index for ranking of scientists really work? Scientometrics 65, 391–392.

Bornmann, L., Daniel, H.D., 2009. The state of h index research: is the h index the ideal way to measure research performance? EMBO reports 10, 2–6.

Costas, R., Bordons, M., 2007. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of informetrics 1, 193–203.

Ding, J., Liu, C., Kandonga, G.A., 2020. Exploring the limitations of the h-index and h-type indexes in measuring the research performance of authors. Scientometrics 122, 1303–1322.

Egghe, L., 2006. Theory and practise of the g-index. Scientometrics 69, 131–152.

Egghe, L., 2010. The hirsch index and related impact measures. Annual review of information science and technology 44, 65–114. Egghe, L., 2011. A disadvantage of h-type indices for comparing the citation impact of two researchers. Research Evaluation 20, 341–346.

Hirsch, J.E., 2005. An index to quantify an individual's scientific research output. Proceedings of the National academy of sciences 102, 16569– 16572.

Hirsch, J.E., 2019. h α: An index to quantify an individual's scientific leadership. Scientometrics 118, 673–686.
Jin, B., Liang, L., Rousseau, R., Egghe, L., 2007. The r-and ar-indices: Complementing the h-index. Chinese science bulletin 52, 855–863.
Khurana, P., Sharma, K., 2022. Impact of h-index on author's rankings: an improvement to the h-index for lower-ranked authors. Scientometrics 127, 4483–4498.
Liu, Y., Yang, Y., 2014. Empirical study of l-sequence: The basic h-index sequence for cumulative publications with consideration of the yearly citation performance. Journal of Informetrics 8, 478–485.
Penfield, T., Baker, M.J., Scoble, R., Wykes, M.C., 2014. Assessment, evaluations, and definitions of research impact: A review. Research evaluation 23, 21–32.
Prathap, G., 2010. The 100 most prolific economists using the p-index. Scientometrics 84, 167–172.
Reed, M., Ferre, M., Martin-Ortega, J., Blanche, R., Lawford-Rolfe, R., Dallimer, M., Holden, J., 2021. Evaluating impact from research: A methodological framework. Research Policy 50, 104147.
Rons, N., Amez, L., 2009. Impact vitality: an indicator based on citing publications in search of excellent scientists. Research Evaluation 18, 233–241.
de Saint-Georges, M., de la Potterie, B.v.P., 2013. A quality index for patent systems. Research Policy 42, 704–719.
Schreiber, M., 2008a. A modification of the h-index: The hm-index accounts for multi-authored manuscripts. Journal of Informetrics 2, 211–216. Schreiber, M., 2008b. To share the fame in a fair way, hm modifies h for multi-authored manuscripts. New Journal of Physics 10, 040201.
Schubert, A., Schubert, G., 2019. All along the h-index-related literature: a guided tour, in: Springer handbook of science and technology indicators. Springer, pp. 301–334.
Todeschini, R., Baccini, A., 2016. Handbook of bibliometric indicators: Quantitative tools for studying and evaluating research. John Wiley & Sons.
Van Leeuwen, T., 2008. Testing the validity of the hirsch-index for research assessment purposes. Research Evaluation 17, 157–160. Zhang, C.T., 2009. The e-index, complementing the h-index for excess citations. PLoS One 4, e5429.
Zhang, C.T., 2013. The h'-index, effectively improving the h-index based on the citation distribution. PloS one 8, e59912.

https://doi.org/10.51408/issi2025_071

Model Construction and Empirical Research of China's Science Structure and Science Development

Qianfei Tian[1], Yunwei Chen[2], Zhiqiang Zhang[3]

[1]*tqf@clas.ac.cn*, [2]*chenyw@clas.ac.cn*, [3]*zhangzq@clas.ac.cn*
National Science Library (Chengdu), Chinese Academy of Sciences, Chengdu 610299 (China)
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190 (China)

Abstract

Science structure is defined as the organic structure formed by the long-term development and change of scientific knowledge. In addition to the structure of the global scientific network, each country has its own national science structure. We firstly reviewed representative research on science structure from different fields. Secondly, we constructed a model of science structure at the national level from four dimensions focusing on the research field of scientometrics. Thirdly, empirical research was carried out using more than 40 years of literature data, revealing the development and growth trend of China's science. Finally, the role of China's science in the world science development and its position in global scientific collaboration were observed, and brief suggestions were provided for the development of science in China.

Introduction

With the rapid development of science, the structure of science is constantly evolving. Based on the relevant research of Kuhn(1962), Wei Junchao (2011), Li Jie (2016), Zhang Ruihong and Chen Yunwei (2019) and other scholars, science structure is defined as an organic structure formed by the long-term development and change of scientific knowledge, which is not subject to one's will. It can reflect the logical relationship of science as a whole, and the knowledge structure of a single research field.

How to objectively quantify or study the evolution of science structure, deeply observe and summarize its evolution laws and characteristics, and lay the foundation for the efficient and high-quality development of science has become one of the important topics studied and discussed by many philosophers, information scientists, economists, et al. (Tian Q, Chen Y, Zhang Z, 2024). According to Fortunato S, Bergstrom C, Borner K (2018), science can be described as a complex, self-organizing, and evolving multiscale network. Science is multi-dimensional, requiring the analysis of the scientific performance of individuals, teams and countries from multiple dimensions (Vinkler P, 2010).

From the perspective of scientific networks, there is only one global scientific network. Outside of the global network, each country has its own national science system (Wagner C S, Park H W, Leydesdorff L, 2015). In order to observe the development and evolution of scientific models, we can research from multiple dimensions such as time (e.g., decade, year or month span), space (e.g., global, China, reference country), research field (e.g. subject, discipline, research area), collaborators (e.g., collaboration country, collaboration institution) and so on

(Scharnhorst A, Börner K, Besselaar P, 2012). This paper aims to construct science structure model at the national level, focus on the development and growth trend of China's science (including the development and evolution trend of major fields of science in China), and observe the role of China's science in world science development and its position in global scientific collaboration.

Literature review of science structure

The concept of science structure has been studied by scholars in many fields such as scientific philosophy, scientometrics, and scientific economics, which respectively affirmed the existence of science structures and constructed some models, quantified global science structure and discipline layout in a country, and expanded the research content of science structure to focus on scientific efficiency.

Science structure research from philosophy of science

One of the basic tasks of philosophy of science is to comprehensively reveal the structure, functional transformation and scientific development laws of the entire human science (especially the modern scientific system) (Liu B, Deng P, 1989). After researchers affirmed the existence of science structures (Shen X, Liu S, Zhao H, 1981), they continued to construct the model/levels of science structure.
Leydesdorff (2001) proposed a multi-dimensional scheme to describe "world of science"(Figure 1).

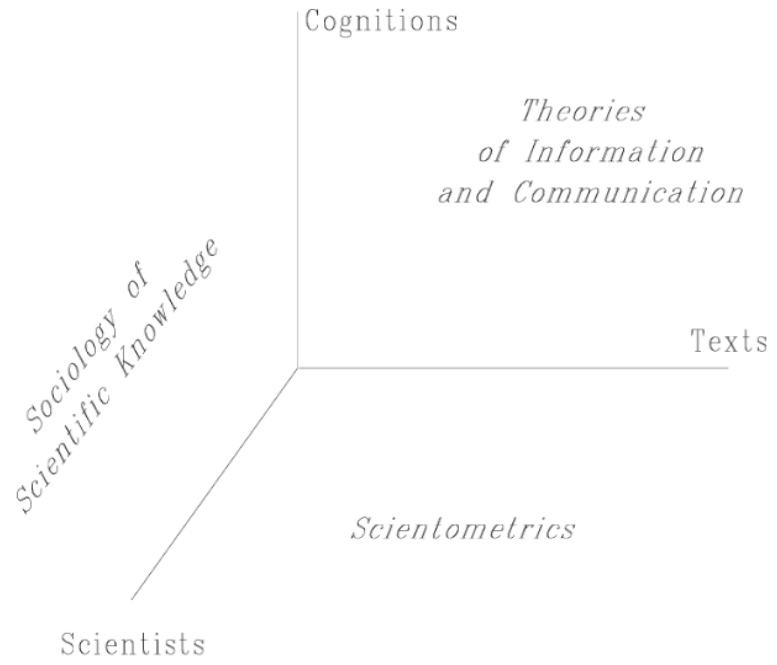


Figure 1. Study of the science as a multidimensional problem.

Thomas R. Blackburn (1973) described science structure as 3 levels, including level of material structure (scientific institutes, material conditions for scientific

work, etc.), level of social structure (scientists, social networks) and level of intellectual structure (scientific knowledge, scientific research).

Science structure research from scientometrics

In the field of scientometrics, researchers investigated ways (Table 1) to quantify global science structure and discipline layout in a country, to reveal global collaboration network and its evolution, to analyze knowledge units such as keywords and themes.

Table 1. Science structure representative research from scientometrics.

Representative researchers (Publishing Year)	*Research aims*	*Years under investigation*	*Main contents or conclusions*
Zhao Hongzhou (1990)	to quantify global science structure and discipline layout in a country	1981-1985	USA, West Europe, Japan etc., their structures of subject become "Polarized" ones, focusing life science; In other, USSR, East Europe, etc. has a "tripartite" science structure, basing on biology, physics and chemistry.
GLÄNZEL (2008)		1991-2005	China joined the triad formed by the USA, EU and Japan, and has transformed the triad into a tetrad.
LI Ning (2019)		1996-2015	China differs significantly from the world's major nations in their research output distributions. China has constantly been comparatively strong in all major fields of physical sciences but weak in areas of life, health, and social sciences.
LIU Yun (2001)	to reveal global collaboration network and its evolution	1994-1998	Systemically measured and evaluated the situation of international collaborating of Chinese basic research from six aspects.
ZHOU Ping (2010)		1997-2007	The authors analyze the dynamics and the national characteristics of China's co-operation in a global context.

			They also study research profile and citation impact of international collaboration with respect to the corresponding domestic 'standards'.
LIU Chengliang (2017)		2014	International scientific collaboration network presents a core-periphery structure with hierarchies, which is composed of 13 core countries and the periphery of 198 countries.
Jyoti Dua (2023)		2000-2020	USA, Germany, England, and China remain the top collaborating partners of India in terms of volume of papers, however, the relative intensity of collaboration with South Korea and Saudi Arabia has increased significantly.
GE Fei (2012)		---	Several principal research methods of science structure and evolution are introduced, including the method of citation analysis, the method of content words analysis and the method of bibliometric combining with content words analysis. The authors suggest that the hybrid method can be applied in researching science structure and evolution and detecting the emerging trends.
LU Wanhui (2019)	to analyze knowledge units such as keywords and themes	---	This paper discussed the application and challenges of knowledge network mining technology in the fields of knowledge organization and management, the construction of scientific knowledge map and the monitoring of discipline development situation by combing the related research of knowledge network concept and type, characteristics and performance, evolutionary

WANG Xiaomei (2024)	2016-2021	analysis methods and indicators. The highly cited papers and 12,620 research frontiers were extracted, and 1,389 research areas were obtained through co-cited cluster analysis, forming a global perspective of science structure map, visually showing the macro structure of scientific research and its internal relationships.

Science structure research from scientific economics

Chinese scholar Gu Xingrong (2006) innovatively proposed that the fundamental task of science and technology was to use scientific and technological progress to offset the marginal rate of diminishing returns in economics. On the basis of the input-output relationship in the economic field, he proposed the structure of "three stations and two transformations" of scientific and technological input-output shown in Figure 2.

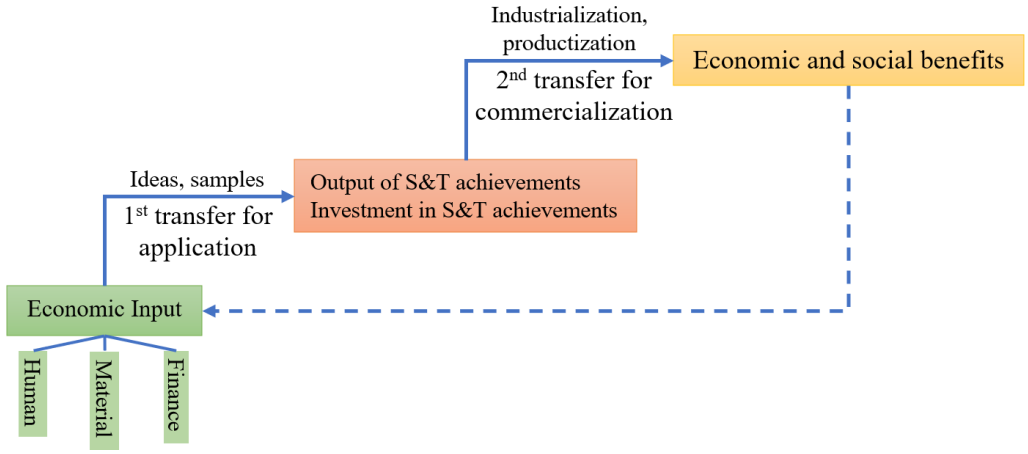


Figure 2. Structure of "three stations and two transformations" of scientific and technological input-output.

May R M (1997) proposed that comparison of scientific output relative to government money spent on research and development (R&D) might be the best measure of the cost effectiveness of spending in support of basic and strategic research. He came to the conclusion of countries' scientific productivity rank descending as: Great Britain, Switzerland, Denmark, Sweden, France, Italy and Germany.

Model construction of China's science structure and its development

Research papers and related changes can reflect how science is organized at an aggregated level (Yang T, 1984). Based on the qualitative and quantitative research of information scientists and scientometrics researchers on the scientific structure and national science, we focused on the research field of scientometrics and constructed a model of the development and evolution of China's science structure in 4 dimensions, including science productivity, science impact, science equilibrium, and science collaboration. The specific dimensions and indicators are listed in Table 2. In terms of time scale, it includes not only long-term annually data and summary data for 42 years from 1980 to 2021, but also evolutionary data for 4 consecutive decades (1980-1989, 1990-1999, 2000-2009, and 2010-2019). At the spatial scale, in addition to focusing on China, it also includes global or USA data for reasonable comparison, which not only reveals the development and internal logic of China's own science structure, but also better presents China's position in the global scientific network.

Table 2. The 4 dimensional model for development and evolution of China's science structure.

Dimensions	*Indicators*	*Time Scale*
Productivity of China's science	Number of international papers per year, Global share (%), annual growth rate (%)	42 years (annually)
Impact of China's science	Number of citations per paper, Number of top 1% cited papers, Global share (%); (and compared with the corresponding data in the USA) Category Normalized Citation Impact, CNCI; (and compared with the corresponding data in the USA)	42 years (annually)
Equilibrium of China's science	Number of international papers in each field of science, Global share (%), Revealed Comparative Advantage, RCA across major fields of science; CNCI each field of science (and decade evolution); Weight and polarization degree of each field of science (and decade evolution)	42 years (annually) 4 decades
Collaboration of China's science	Number of international collaborative papers, Share of China's total paper (%), Global share (%); Top 10 collaboration countries (and decade evolution); Collaboration networks and evolution (and compared with the corresponding data in the USA)	42 years (annually) 4 decades

Data source and method

Data was downloaded from the InCites platform of Clarivate Analytics Web of Science (WOS) database, including the annual number of papers (limited to article and review) from 1980 to 2021 in China (not including data from Hong Kong, Macao and Taiwan) and the world, the number of papers in 22 fields of science according to the Essential Science Indicators (ESI) categories, and the number of citations. In addition to the aggregate analysis from 1980 to 2021, the comparative analysis of data from 4 consecutive decades (1980 to 1989, 1990 to 1999, 2000 to 2009, and 2010 to 2019) was also carried out. In order to analyze China's position and its evolution in the global science collaboration network, the Science Citation Index (SCI) and Social Science Citation Index (SSCI) in the WOS core database (excluding data from Hong Kong, Macao and Taiwan) were used to retrieve the data of international scientific research collaboration. Full counting method was adopted, which was the main choice in most national bibliometric studies (Chen L, Yang L, Ding J, 2018; Braun, T, Glänzel, W, & Schubert, A, 2005; Leydesdorff L, 1988).

Empirical research of China's science structure and its development

Research dimension in productivity of China's science

Figure 3 plots the number of China's annual papers from WOS, 1980 to 2021, and the ratio of China's papers to the global total. The absolute quantity of China's scientific papers and relative ratio in global total papers have been increasing. In the past 40 years, the number of China's WOS papers has increased exponentially, which can be shown by the Exponential Trendline in figure 3.

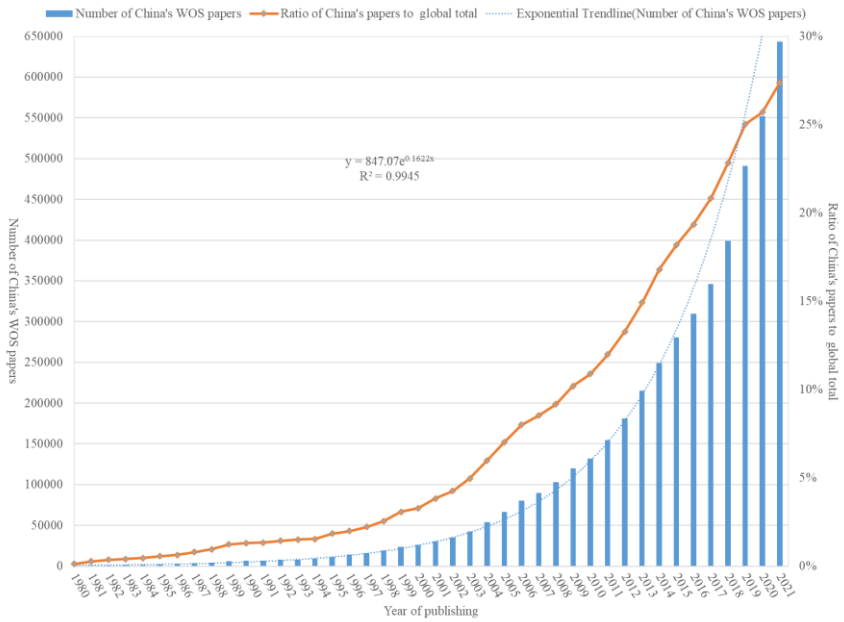


Figure 3. Time series trend of number of China's paper in WOS (1980 to 2021).

In the 21st century, the number of China's WOS papers has increased rapidly year by year, and the corresponding ratio of the global total has also increased rapidly. In 2008, the number of China's WOS papers exceeded 100,000, accounting for 9.16% of the world's total. In 2013, it exceeded 200,000 articles, accounting for 14.92% of the world's total; in 2016, it exceeded 300,000 articles, accounting for 19.35% of the world's total; nearly 400,000 in 2018; nearly 500,000 in 2019; in 2021, it exceeded 640,000 articles, accounting for 27.38% of the world's total. According to Mitutomo Y (1963) (a professor and a historian of science in Japan), who defined "center of scientific activity": a country whose scientific achievements account for more than 25% of the world's total, China can be regarded as one of the "world's scientific center of WOS papers" since 2019.

Research dimension in impact of China's science

In order to analyze the quality level and influence of scientific papers in China, this part observes and discusses the average citation frequency of papers, the ratio of the top 1% cited papers to the corresponding global value, and the Category Normalized Citation Impact (CNCI) from InCites platform.

(1) Citations per paper and top 1% cited papers

As shown in Figure 4, as the proportion of China's WOS papers to the world's total has increased year by year, the proportion of China's WOS papers citations to the world's total has also increased yearly, but the latter has been lower than the former from 1980 to 2013. In 2019, the proportion of China's WOS papers citations exceeded the corresponding value of the USA. From the perspective of the proportion of the top 1% cited papers to the world, the proportion of China's top 1% of cited papers to the world has been significantly lower than that of the USA for a long time. The proportion of China's top 1% cited papers has exceeded 20% since 2015, while the corresponding percentage of the USA has fallen below 50%. The gap has narrowed significantly, until 2020, the proportion of China's top 1% cited papers to the world has exceeded the corresponding proportion of the USA.

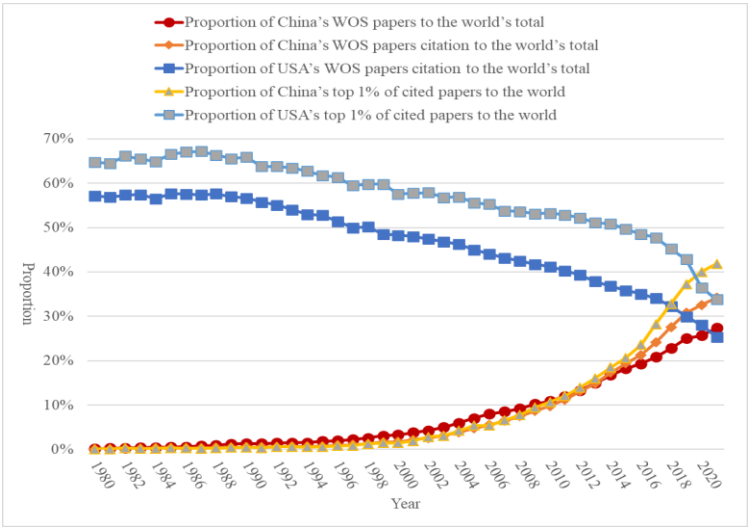


Figure 4. The proportion of China's and USA's top 1% cited papers to the world.

(2) CNCI and annual comparison between China and USA

The Category Normalized Citation Impact (CNCI) (He C, Li W, 2022; Incites, 2025) of a document is calculated by dividing the actual count of citing items by the expected citation rate for documents with the same document type, year of publication and subject area. When a document is assigned to more than one subject area an average of the ratios of the actual to expected citations is used. The CNCI of a set of documents, for example the collected works of an individual, institution or country/region, is the average of the CNCI values for all the documents in the set. CNCI solves the problem of incomparability between different countries, years and fields of science. The world average is 1, and if the CNCI value is greater than 1, it means that the influence of the paper exceeds the world average.

For a single paper that is only assigned to one subject area, this can be represented as:

$$CNCI = \frac{c}{e_{ftd}}$$

For a group of papers, the CNCI value is the average of the values for each of the papers:

$$CNCI_i = \frac{\sum_i CNCI_{each\ paper}}{p_i}$$

Equation Key:
e: Expected citation rate or baseline;
c: Times cited;
p: Number of papers;
f: The field or subject area;
t: Year;
d: Document Type;
i: Entity being evaluated (institution, country/region, person, etc.)

From the perspective of the ratio of the top 1% cited papers to domestic papers in Figure 5, the percentage of China (circular markers) is continually lower than that of the USA (diamond markers). From the perspective of the influence of CNCI, this value in China (dotted column) was also lower than that of the USA (solid column), until 2020 it slightly exceeded but fell back in 2021. Judging from the trend of China's CNCI, the value continued to grow, surpassing the world average for the 1st time in 2012 at 1.021.

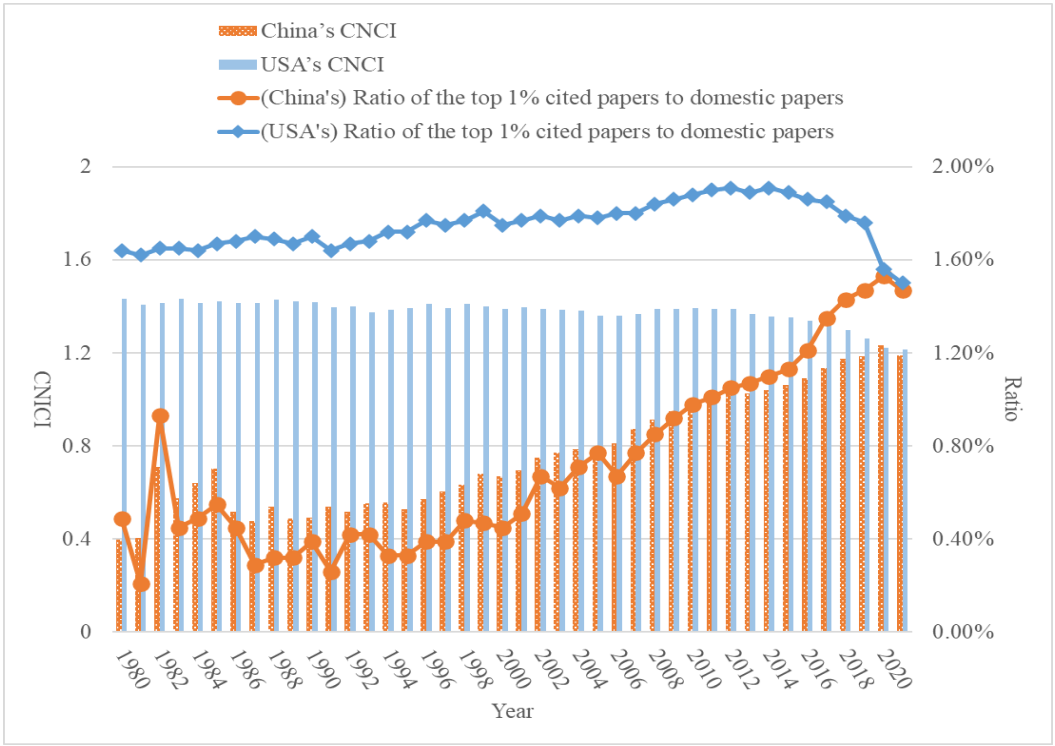


Figure 5. CNCI in China, USA and percentage of TOP 1% Citations in each country.

Research dimension in equilibrium of China's science

(1) The absolute value and Revealed Comparative Advantage of China's percentage of various fields of science

According to the absolute value of the percentage of China's each field of science from 1980 to 2021, the top five dominant fields of science in China are: chemistry, engineering, materials sciences, clinical medicine and physics. The percentages of these fields of science to the total number of China's papers are 17.561%, 13.694%, 11.610%, 10.199% and 9.608%, respectively. The corresponding percentages of other fields of science are less than 5%.

According to the difference between the percentage of China's each field of science and corresponding value of the world's from 1980 to 2021, China has 9 fields of science with a numerical advantage relative to the world (on the left side of the dotted line in Figure 6), and the top 5 fields of science are: materials sciences, chemistry, engineering, physics and computer science, which are 6.346%, 5.927%, 5.245%,

1.624% and 1.179%, higher than the corresponding global percentage, respectively. Although clinical medicine ranks the 4th in China in terms of absolute percentage, it ranks last in terms of percentage difference with the world, 8% lower than the corresponding global percentage.

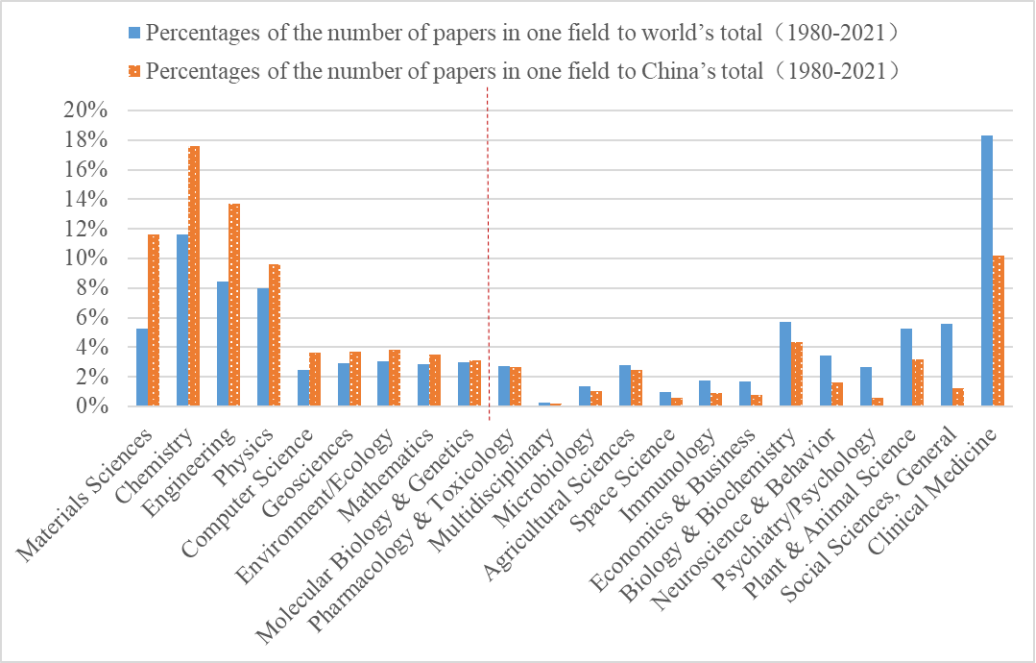


Figure 6. Comparative chart of China's and the world's field of science percentages (in descending order by the difference) (1980 to 2021).

Compared with the development of various fields of science in the world, what is the competitive advantage of China's each field of science? The Revealed Comparative Advantage (RCA) of China's each field of science compared with the world can be calculated as: RCA = the percentage of a certain field to China's total / the percentage of a certain field of science to the world's total. If the RCA value is greater than 1, this field in China has a significant comparative advantage to the world.

According to table 3, shown in descending order of RCA of China's various fields of science from 1980 to 2021, there are 9 fields with RCA values greater than 1 (background filled), namely: materials sciences, engineering, chemistry, computer science, geosciences, environment/ecology, mathematics, physics, and molecular biology & genetics.

According to the RCA values (**in Bold**) of the 4 consecutive decades, there are a total of 6 China's fields (with underline) that show the comparative advantage of numerical explicitness of papers, namely: materials sciences, engineering, chemistry, geosciences, mathematics and physics.

Table 3. RCA value and evolutionary dynamics of China's various fields of science.

Fields of science & RCA	*1980-2021*	*1980-1989*	*1990-1999*	*2000-2009*	*2010-2019*
Materials Sciences	**2.21**	1.51	2.47	2.53	1.86
Engineering	**1.62**	1.53	1.32	1.22	1.46
Chemistry	**1.51**	1.52	2.22	2.05	1.49
Computer Science	**1.48**	1.00	0.67	1.09	1.42
Geosciences	**1.26**	1.40	1.03	1.19	1.21
Environment/Ecology	**1.24**	0.55	0.53	0.85	1.05
Mathematics	**1.23**	2.92	2.63	1.73	1.15
Physics	**1.20**	2.80	2.14	1.67	1.30
Molecular Biology & Genetics	**1.05**	0.28	0.24	0.53	1.14
Pharmacology & Toxicology	**0.98**	1.34	0.77	0.79	1.01
Agricultural Sciences	**0.87**	0.34	0.35	0.59	0.84
Biology & Biochemistry	**0.77**	0.25	0.50	0.69	0.96
Microbiology	**0.75**	0.34	0.27	0.56	0.80
Multidisciplinary	**0.66**	1.89	5.88	0.68	0.82
Plant & Animal Science	**0.61**	0.36	0.46	0.61	0.67
Space Science	**0.61**	1.48	0.96	0.87	0.65
Clinical Medicine	**0.56**	0.61	0.29	0.32	0.60
Immunology	**0.52**	0.23	0.15	0.35	0.56
Neuroscience & Behavior	**0.48**	0.27	0.23	0.33	0.54
Economics & Business	**0.44**	0.21	0.17	0.19	0.41
Psychiatry/Psychology	**0.22**	0.10	0.11	0.09	0.20
Social Sciences, General	**0.22**	0.39	0.21	0.13	0.19

(2) Decade evolution of CNCI in China's various fields of science

From the perspective of longitudinal temporal evolution (Figure 7), from the 80s of the 20th century (triangle markers) to the 20s of the 21st century (diamond markers), the number of CNCI higher than 1 in China's various fields of science increased from 2 fields to 14 fields.

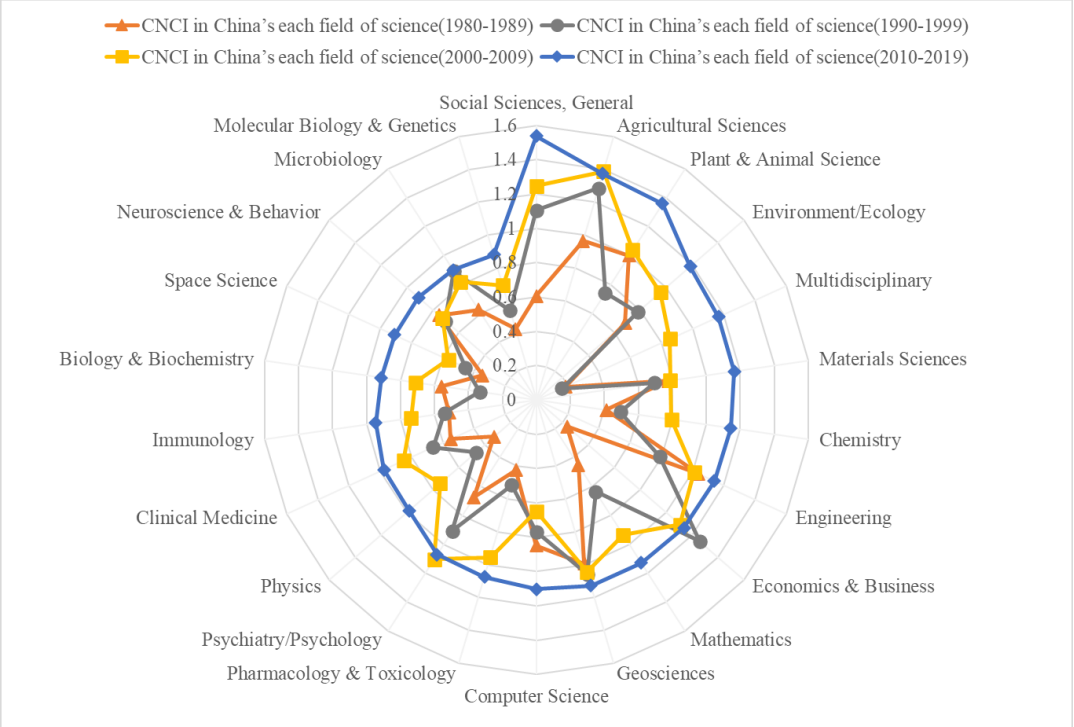


Figure 7. Dynamic of CNCI value evolution in China's various fields.

(3) Decade evolution of China's field weights and polarization degree

Zhao H (1990) defined indicators such as "field weight" and "field polarization degree" to measure the status and influence of a particular field in the overall science structure of a country. **The weight p** of a field in China and **the degree** α of polarization of the structure of a field in China can be calculated by the following two formulas.

$$p=\sqrt{a^2+b^2}$$

Note: a is the weight of China's each field compared to the global, and b is the weight of China's each field compared to China's all fields.

$$\alpha=1-\frac{p_{min}}{p_{max}}$$

Table 4 below lists the weights of China's 22 fields in multiple periods, the and the top three belong to materials science, chemistry and engineering.

Table 4. China's field weights in various periods.

Fields and their weights	*Various periods*				
	1980-2021	*1980-1989*	*1990-1999*	*2000-2009*	*2010-2019*
Materials Sciences	**0.308**	0.038	0.107	0.225	0.379
Chemistry	**0.263**	0.168	0.259	0.301	0.338
Engineering	**0.251**	0.094	0.099	0.128	0.314
Computer Science	**0.196**	0.017	0.019	0.086	0.277
Physics	**0.183**	0.243	0.217	0.199	0.269
Geosciences	**0.167**	0.042	0.036	0.094	0.237

Environment/Ecology	**0.166**	0.011	0.016	0.067	0.207
Mathematics	**0.163**	0.091	0.092	0.137	0.225
Molecular Biology & Genetics	**0.139**	0.008	0.009	0.042	0.223
Pharmacology & Toxicology	**0.130**	0.039	0.026	0.062	0.198
Clinical Medicine	**0.125**	0.121	0.054	0.064	0.159
Agricultural Sciences	**0.115**	0.010	0.011	0.047	0.165
Biology & Biochemistry	**0.109**	0.021	0.037	0.065	0.192
Microbiology	**0.098**	0.005	0.007	0.042	0.156
Plant & Animal Science	**0.085**	0.023	0.028	0.056	0.134
Multidisciplinary	**0.085**	0.022	0.127	0.050	0.159
Space Science	**0.079**	0.021	0.024	0.065	0.126
Immunology	**0.068**	0.004	0.004	0.026	0.110
Neuroscience & Behavior	**0.064**	0.009	0.010	0.027	0.106
Economics & Business	**0.058**	0.004	0.004	0.015	0.080
Social Sciences, General	**0.031**	0.024	0.011	0.011	0.038
Psychiatry/Psychology	**0.030**	0.003	0.004	0.007	0.040

Using the formula of calculating the polarization degree of a country's field structure, table 5 lists polarization degrees of fields in China and the USA in different periods. The value ranges from 0 to 1. The smaller the value, the more balanced field structure a certain country has. From 1980 to 2021, China's field polarization degree is higher than that of the USA. From the perspective of the longitudinal sequence of the 4 decades, the polarization degrees of China's fields shows a decreasing trend, indicating that the balance of field structure and layout has been improved.

Table 5. The field polarization degree and dynamic evolution of China and USA.

Field polarization degree	*Various periods*				
	1980-2021	*1980-1989*	*1990-1999*	*2000-2009*	*2010-2019*
China	**0.904**	0.987	0.986	0.978	0.900

Research dimension in collaboration of China's science

(1) Scale and trend of China's science collaboration

Judging from the number of international collaborations in China shown in Figure 8, the number of international collaboration papers in China shows a sustained and exponential growth pattern. With the advent and development of the era of big science, the proportion of global collaborative papers (circular markers) has increased year by year, and the proportion of China's collaborative papers to the world's collaborative papers (dotted line) is no exception. The proportion of USA collaborative papers to the world's collaborative papers (solid line) exceeded 50% from 1980 to 1989, but has been declining year by year since then.

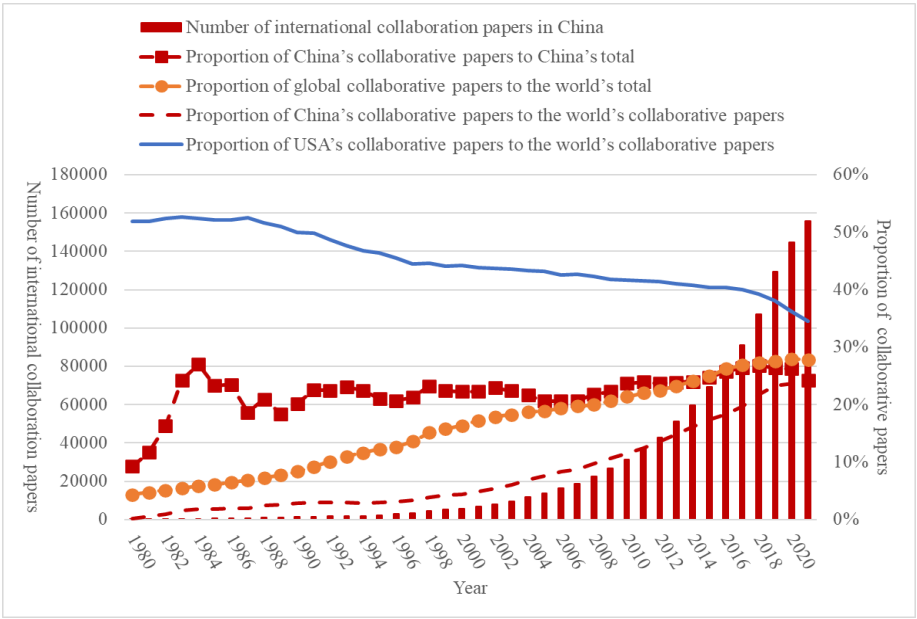


Figure 8. Number of WOS collaborative papers in China and collaboration proportion of China, USA (1980-2021).

From the perspective of China's collaboration rate (red dotted line), from 1980 to 1984 it was a rapid climbing stage, from 9.27% to 27.01%, and then a slight fluctuation trend of declining-growing was repeated. The collaboration percentages in the past 4 decades were 20.38%, 22.01%, 21.60% and 25.35% respectively.

Table 6 shows the changes in the number and proportion of China's international collaboration papers in the past 4 decades. With the number of global collaboration papers and the percentage of global collaboration continue to rise, China's international collaboration rate (the proportion of China's international collaboration papers to China's total) has remained between 20%~26% in the 4 decades. The number of cooperative publications has increased from more than 5,000 to nearly 700,000, an increase of about 139 times; China's collaboration percentage of global collaboration total has increased from 2.06% to 18.34%, and the percentage doubles almost every decade. On the whole, China's international scientific research collaboration is becoming more and more active.

Table 6. Number and related proportion of international collaboration papers in China and globally.

Number and proportion	*4 decade series*			
	1980-1989	*1990-1999*	*2000-2009*	*2010-2019*
Number of international collaboration papers in China	5052	26922	139885	699227
Proportion of international collaboration papers in China/%	**20.38**	**22.01**	**21.60**	**25.35**

Number of international collaboration papers	245674	767897	1785878	3813275
Proportion of international collaboration papers /%	6.12	12.31	18.86	24.91
Proportion of China's collaborative papers to the world's collaborative papers /%	**2.06**	**3.51**	**7.83**	**18.34**

(2) Countries distribution of China's collaborative papers

From 1980 to 2021, China carried out international science collaboration with more than 200 countries/regions, with a total of 1,171,904 international collaboration papers in China. The top 10 collaborative countries are shown in Figure 9.

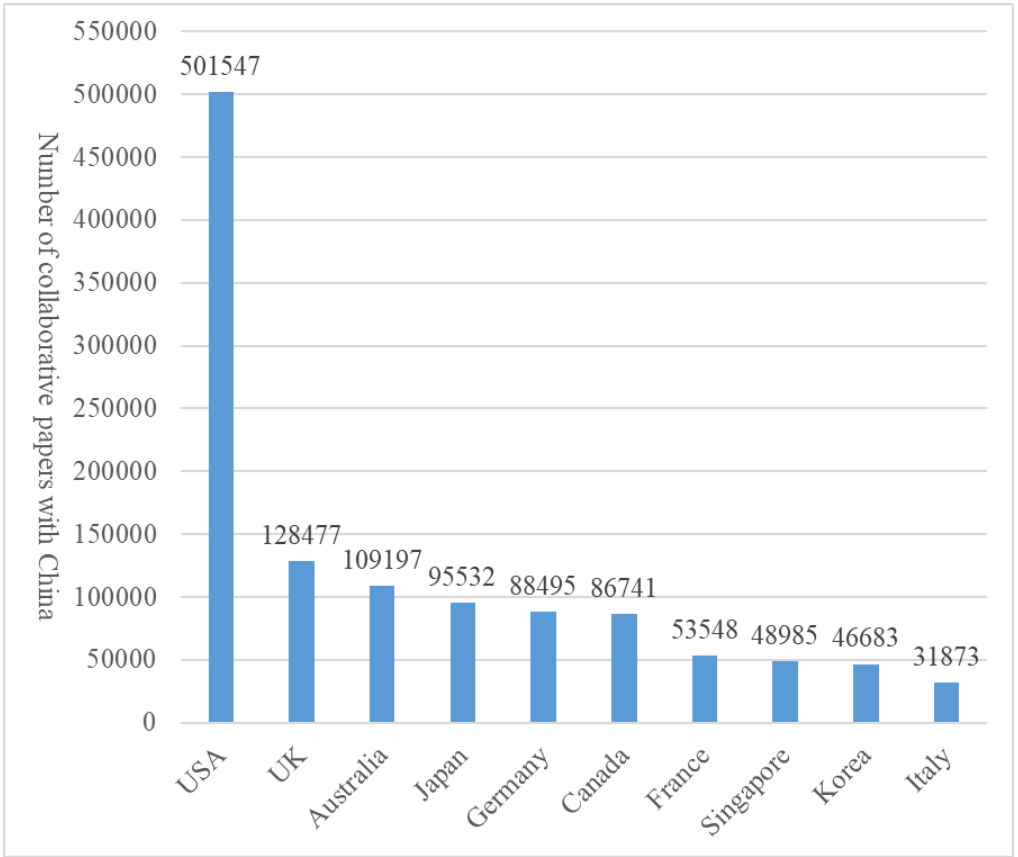


Figure 9. Top 10 countries and numbers of collaborative papers with China (1980-2021).

From the perspective of the evolution of top 10 collabrative countries and the number of papers in the 4 decades (Table 7), the USA occupied the first position in the number of collabrative papers with China for 4 consecutive decades, and Japan occupied the second position in the first 3 decades, and then gave way to the United Kingdom in the 4th decade (2010-2019).

Table 7. Top 10 countries and number of papers in collaboration with China (4 decade series).

Rank	*Country (number)*			
	1980-1989	*1990-1999*	*2000-2009*	*2010-2019*
1	USA (2702)	USA (9831)	USA (54442)	USA (322066)
2	Japan (567)	Japan (4380)	Japan (21038)	UK (74633)
3	UK (421)	Germany (3062)	UK (12912)	Australia (66735)
4	Canada (400)	UK (2694)	Germany (12630)	Canada (52046)
5	Germany (370)	Canada (2010)	Canada (9930)	Japan (51412)
6	France (297)	France (1682)	Australia (9026)	Germany (50762)
7	Australia (196)	Italy (1327)	France (7154)	France (31728)
8	Italy (145)	Australia (1250)	South Korea (6375)	Singapore (29566)
9	Sweden (117)	Netherlands (797)	Singapore (5935)	South Korea (27608)
10	Swiss (92)	South Korea (730)	Sweden (3419)	Netherlands (18368)

Following North American countries (the USA, Canada), European countries (the United Kingdom, Germany, France, Italy, Sweden, Switzerland), Australia in Oceania, and Japan in Asia, the Netherlands in Europe, South Korea and Singapore in Asia were also among the top 10 partner countries in China (Table 8). In the 2nd decade (1990-1999), the Netherlands (with underline) and South Korea (with underline) ranked among the top 10 collabrative countries. In the 3rd decade (2000-2009), Singapore (with underline) ranked 9th in the top 10 of China's collabrative countries.

Table 8. Top 10 collaboration countries with China and decadal evolution.

Top 10 collaboration countries with China	*4 decade series*			
	1980-1989	*1990-1999*	*2000-2009*	*2010-2019*
USA	√	√	√	√
Japan	√	√	√	√
UK	√	√	√	√
Canada	√	√	√	√

Germany	√	√	√	√
France	√	√	√	√
Australia	√	√	√	√
Italy	√	√		
Sweden	√		√	
Swiss	√			
Netherlands		√		√
South Korea		√	√	√
Singapore			√	√

(3) Changes in the pattern of China's global collaboration network

In order to reveal China's global collaboration network more clearly and compare it with the USA, Table 9 lists the node characteristics of China and the USA in the global scientific paper collaboration network in the past 4 decades (when calculating the international scientific research collaboration network, the edge of the collaboration relationship below 40 is removed). In graph theory and network analysis, Centrality is a metric to judge the importance/influence of nodes in a network. Degree centrality refers to the number of connections a node has. Betweenness centrality is defined in terms of the proportion of shortest paths that go through a node for each pair of nodes. Closeness centrality is the inverse of the sum of the shortest path lengths between a node and all other nodes in the network. Eigenvector centrality is related to the centrality of adjacent nodes of a node and it assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes.

The USA has been at the heart of the network for the past 4 decades. China's degree, weighted degree, closeness centrality, betweeness centrality and eigen centrality in the network have all shown a monotonous growth trend in the 4 decades, and the gap between China and the corresponding values of the USA is gradually narrowing.

Table 9. Node characteristics of China, USA in global scientific collaboration network (4 decade series).

Node characteristics	*China*				*USA*			
	1980-1989	*1990-1999*	*2000-2009*	*2010-2019*	*1980-1989*	*1990-1999*	*2000-2009*	*2010-2019*
degree	14	47	73	**145**	84	122	154	**181**
weighted degree	13130	78656	368802	**2105508**	333156	927650	2100290	**5012286**
closeness centrality	0.498	0.595	0.639	**0.801**	0.796	0.855	0.923	**0.941**
betweeness centrality	0	30.603	231.177	**367.344**	2480.888	3005.583	3994.444	**2458.131**
eigen centrality	0.432	0.732	0.799	**0.966**	1	1	1	**1**

Conclusion of 4 decades evolution characteristics of China's science structure

From the empirical research and multi-dimensional analysis of the development process of China's science structure, one may conclude the development and evolution of China science structure in the past 4 decades as "starting-consolidating-improving-rising". Each key indicator selected from the 4 dimensions of productivity, impact, equilibrium and collaboration, could be used to show the development trends in the 4 decades (Table 10).

Table 10. Key indicators and conclusions of the development and evolution of China's science structure.

Dimension: Key indicator	*starting*	*consolidating*	*improving*	*rising*	*Key items*
	1980-1989	*1990-1999*	*2000-2009*	*2010-2019*	
Productivity: ratio of China's WOS papers to the global total	0.62%	1.96%	6.84%	18.02%	China can be regarded as one of the "world's scientific center of WOS papers" since 2019
Impact: CNCI	0.54	0.59	0.84	1.10	China's CNCI value continues growing, surpassed the world average for the 1st time in 2012 at 1.021; In 2020 China's CNCI slightly exceeded USA's but fell back in 2021.
Equilibrium: field polarization degree	0.987	0.986	0.978	0.900	Balance of field structure and layout in China has been improved.
Collaboration: number of countries collaborated more than 40 papers with China	14	47	73	145	More and more countries are collaborating with China.

The development of China's science structure, which is shown by the four dimensions of productivity, impact, equilibrium and collaboration, has continued to improve, especially in the past decade, and important breakthroughs have been made in the dimension of productivity. On the basis of the continuous expansion of the scale of scientific output, China's research field structure has been continuously improved.

But there are still some challenges in China's scientific development. First, although the scale of scientific output has made leaps and breakthroughs, basic science and technological breakthroughs still require long-term accumulation and resource investment. Second, although China has grown into a major country in scientific scale, the "qualitative" change and breakthrough has not yet been fully realized. Third, from the perspective of field structure, the overall equilibrium of China's fields structure is still obviously insufficient. Last, the scale of China's global scientific cooperation is expanding, but the gap is still large compared with the USA, which occupies the core position of the network.

In the future, China could continue to optimize its science structure, starting from

various aspects such as scientific research investment, high-quality development, field layout, and international scientific collaboration, to promote the development of scientific undertakings, and achieve the goal of becoming a major science center in the world.

Research limitations and future research directions

This study has the following limitations, and future research can be further improved.
(1) The three theoretical streams cited—philosophy of science, scientometrics, and scientific economics—do not effectively converge. The empirical study only focuses on scientometrics.
(2) The data source is not comprehensive, and only WOS international papers are used to characterize the evolution of China's science structure.
(3) The study focuses on the analysis of the historical characteristics of China's science structure, which lacks future trend prediction and comparison with more countries.
(4) There is a lack of in-depth discussion of the current situation and causes, and a lack of comprehensive evaluation of China's science structure.
In the future, on the basis of existing scientometrics with the scientific papers as the core, more unstructured information and data from the science community such as science and technology policy, scientific research investment related to funding projects, researchers and financial resources data, more detailed disciplinary classification and knowledge data can be considered, so as to understand and explore the China's science structure more comprehensively and concretely.

Acknowledgements

Funded by Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2021171).
We would like to thank Prof. YANG Liying, Assist. Prof. ZHAI Yanqi, and Assoc. Prof. LIU Chunjiang for data acquisition and data processing for this paper. Appreciation to two anonymous experts who gave professional comments and suggestions about the paper.

References

Blackburn T R. (1973). Information and the Ecology of Scholars. *Science*, 181, 1141-1146.
Braun, T., Glänzel, W., & Schubert, A. (2005). Assessing assessments of British science. Some facts and figures to accept or decline. *Scientometrics, 15*, 165-170.
Chen Liyue, Yang Liying, Ding Jielan. (2018). Review on Full Counting Method and Fractional Counting Method in Scientometric Research. *Library and Information Service*, 62(23), 132-141.
Dua, J., Singh, V. K., Lathabai, H. H. (2023). Measuring and characterizing international collaboration patterns in Indian scientific research. Scientometrics, 128(9), 5081-5116.
Fortunato S, Bergstrom C, Borner K, et al. (2018). Science of science. *Science*, 359(6379),eaao0185.
Ge Fei, Tan Zongying. (2012). Review of Science Structure and Evolution of Bibliometric Methods [J]. *Journal of Intelligence*, 31(12), 34-39, 50.
Glänzel W, Debackere K, Meyer M. (2008). 'Triad' or 'Tetrad'? on Global Changes in a Dynamic World[J]. *Scientometrics*, 74(1), 71–88.
Gu Xingrong. (2006). Analysis on Input-output of S&T. *Industrial & Science Tribune*, 1, 84-87.

He Canfei, Li Wentao. (2022). Spatial-temporal Evolution and Driving Forces of China's International Scientific Collaboration Network. *China Soft Science*, 7, 70-81.

InCites help center. (2025). *Category Normalized Citation Impact (CNCI)*. Retrieved Jan 4, 2025 from: https://incites.zendesk.com/hc/en-gb/articles/25087312115601-Category-Normalized-Citation-Impact-CNCI.

Kuhn T S. (1962). *The Structure of Scientific Revolutions: 50th Anniversary Edition*. Chicago: University of Chicago Press.

Leydesdorff L. (2001). *The Challenge of Scientometrics: the Development, Measurement, and Self-organization of Scientific Communications*. Florida: Universal Publishers, 1-25. https://repository.arizona.edu/bitstream/handle/10150/105095/Scientometrics.pdf?sequence=1&isAllowed=y.

Leydesdorff Loet. (1988). Problems with the 'measurement' of national scientific performance. *Science and Public Policy*, 15(3), 149 – 152.

Li Jie. (2016). *Science Structure and Topics Evolution of Safety*. Beijing: Capital University of Economics and Business.

Li Ning. (2019). Disciplinary distribution of China's research outputs: Evolutionary patterns and contributing factors. *Science Research Management*, 40(1), 1-11.

Liu Bo, Deng ping. (1989). A new exploration of the scientific development model. *Studies in Philosophy of Science and Technology*, 1, 44-48,50.

Liu Chengliang, Gui Qinchang, Duan Dezhong, et al. (2017). Structural heterogeneity and proximity mechanism of global scientific collaboration network based on co-authored papers. *Acta Geographica Sinica*, 72(4), 737-752.

Liu Yun, Chang Qing. (2001). Scientometrical measurement and evaluation on international collaboration of basic research in China. *Journal of Management Sciences in China*, 15(1), 64-73.

Lu Wanhui. (2019). Review of the evolution and application of knowledge networks. *Information Studies: Theory & Application*, 42(8), 138-143.

May R M. (1997). The Scientific Wealth of Nations. *Science*, 275, 793-795.

Mitutomo Yuasa. (1963). *Center of scientific activity: its shift from the 16th to the 20th century*. Nagoya University.

Scharnhorst A, Börner K, Besselaar P. (2012). Models of Science Dynamics. Heidelberg: Springer.

Shen Xianjia, Liu Shuzi, Zhao Hongzhou, et al. (1981). Questions on science construction & science. *Science of Science and Management of S&T*, 1, 21-22.

Tian Qianfei, Chen Yunwei, Zhang Zhiqiang. (2024). Scientific Structure Review and Its Hierarchical System Construction. *Journal of the China Society for Scientific and Technical Information*, 43(6), 747 759.

Vinkler P. (2010). *The Evaluation of Research by Scientometric Indicators*. Oxford: Chandos Publishing.

Wagner C S, Park H W, Leydesdorff L. (2015). The Continuing Growth of Global Cooperation Networks in Research: A Conundrum for National Government. *PLoS ONE*, 10(7), e0131816.

Wang Xiaomei, Li Guopeng, Chen Ting. (2024). *Mapping Science Structure*. Retrieved June 20, 2024 from: http://www.casisd.cas.cn/ttxw1/zlyjytt/202405/P020240529314820155762.pdf.

Wei Junchao, Wei Haiyan. (2011). Review on Analysis Methods of Science Structure and Evolution. *Library & Information*, 4, 48-52.

Yang Tingjiao. (1984). Method for micro science structure research – citation analysis. *Technology and Market*, 4, 38-43.

Zhang Ruihong, Chen Yunwei, Deng Yong. (2019). A Review of Community Discovery in Hybrid Network for Science Structure Analysis. *Library and Information Service*, 63(4), 135-141.

Zhao Hongzhou, Jiang Guohua, Zheng Wenyi. (1990). On the Structure of Basic Subjects—the Compare of Two Kinds of Structures of U.S.S.R. & U.S.A.. *Bulletin of National Natural Science Foundation of China*, 2, 41-48.

Zhou P, Glänzel W. (2010). In-depth Analysis on China's International Cooperation in Science. *Scientometrics*, 82, 597-612.

https://doi.org/10.51408/issi2025_010

Network Position Matters: Collaborative Strategies, Talent Mobility, and Exploratory Innovation in Teams

Xiaoling Cheng[1], Jiajie Wang[2], Lele Kang[3]

[1] *xiaolingcheng@smail.nju.edu.cn,* [2] *jiajiewang@smail.nju.edu.cn,* [3] *lelekang@nju.edu.cn*
Laboratory of Data Intelligence and Interdisciplinary Innovation, School of Information Management, Nanjing University No. 163 Xianlin Avenue, Nanjing (China)

Abstract

As innovative teams increasingly depend on external knowledge, talent mobility has emerged as a crucial mechanism for acquiring novel and diversified resources that foster exploratory innovation. Despite this potential advantage, many teams fail to fully leverage newly recruited talents when these individuals lack effective network positions, resulting in underutilized innovative potential. Grounded in the complementary perspectives of collaborative networks and knowledge networks, this study investigates how newly recruited talents' positions in both collaboration and knowledge networks influence teams' exploratory innovation, and examines the interactive effects between these distinct network positions. Drawing from comprehensive data from PATSTAT and COMPUSTAT databases, we identify 65,438 cases of inter-team talent mobility and develop a robust empirical model to test our hypotheses. Our findings reveal that newly recruited talents' collaboration network centrality demonstrates an inverted U-shaped relationship with teams' exploratory innovation—moderate levels of centrality optimize innovation outcomes, while both low and excessively high centrality prove detrimental. Importantly, we discover that higher knowledge network centrality attenuates this curvilinear effect, making the inverted U-shaped curve flatter. This suggests that individuals with extensive knowledge connections maintain relatively stable innovation performance regardless of their collaboration network centrality levels. By elucidating how structural positions across different networks enable newly recruited talents to fully leverage their innovation capacity, this study contributes significant theoretical insights to our understanding of the talent mobility-team innovation link. Additionally, we provide actionable implications for managers seeking to optimize talent deployment strategies and network positioning to maximize exploratory innovation outcomes.

Introduction

As market uncertainty intensifies and innovation competition grows increasingly fierce, recruiting external talents into teams has not only become a key way to acquire novel knowledge and enhance innovative capabilities, but also an essential component of national talent attraction and development strategies (Singh & Agrawal, 2011; Agrawal, McHale, & Oettl, 2017; Wang et al., 2024). An increasing number of innovative team managers and human resources specialists are paying more attention to the relationships among talent selection, cost investment, and the resulting innovation gains. However, the extent to which successfully hired talents can actually generate innovation value for the new team remains a critical challenge for managers (Shi et al., 2023; Song, Almeida, & Wu, 2003; Tandon, Ertug, & Carnabuci, 2020).

The process of talent mobility not only involves the preliminary phases of interviews, background checks, and skills assessments to evaluate the fit between the talent and the team’s needs, but also encompasses the collaboration and integration stage once newly recruited talents join the team (Jain & Huang, 2022). At this stage, newcomers

must collaborate with existing technical members in the team to thoroughly understand the team's existing knowledge and R&D patterns, and subsequently contribute the novel knowledge and experience they have accumulated elsewhere (Wang & Zatzick, 2019). Therefore, if team managers wish to ensure that talent mobility truly promotes innovation, they must pay close attention to both the hiring and integration phases, recognizing that the collaboration strategies employed by newcomers after they enter the team have a direct and critical impact on team innovation.

Previous research on talent mobility has largely focused on questions such as "How to recruit suitable talents" and "How much innovation value do newly recruited talents create for the team" (Wang et al., 2024; Fahrenkopf, Guo, & Argote, 2020; Jain, 2016). Many studies investigate how the social, relational, or knowledge capital of talents influences the process of knowledge transfer (Shi et al., 2023). However, these studies have tended to overlook the integration stage of talent mobility—that is, "How can well-designed collaboration strategies help new recruits adapt to the new innovation environment". In fact, newly recruited talents can only transform their accumulated explicit or tacit knowledge from other teams into new innovative outputs after establishing effective communication and collaboration with the existing members of the new team (Acharya et al., 2022; Zhang, 2021; Myers, 2021; Wang & Zatzick, 2019).

This study focuses on the integration phase of talent mobility. For teams that rely on external knowledge to achieve exploratory innovation, the external experiences and heterogeneous technology sets brought by new talents can substantially drive breakthroughs in new fields and technologies (Song, Almeida, & Wu, 2003; Ge, Huang, & Kankanhalli, 2020; Choudhury, 2017). To elucidate the internal mechanisms by which newcomers' early-stage collaboration strategies affect teams' exploratory innovation, this study integrates network embeddedness theory and exploratory innovation theory (Yang, Lin, & Peng, 2011). This study hypothesizes that the team's exploratory innovation is influenced by these positions, given that newcomers' network locations determine both the quantity and quality of knowledge transfer, as well as the resulting differences in innovation preferences (Bunderson, Van der Vegt, & Sparrowe, 2014). Furthermore, we investigate how newcomers' positions in the knowledge network moderate the above relationship: whereas the collaboration network position reflects social capital, the knowledge network position indicates their embeddedness in terms of knowledge capital (Wang et al., 2014). Combining these two perspectives enables a more comprehensive exploration of how the integration phase of talent mobility affects exploratory innovation.

This study uses the strength of centrality to measure the quality of network positions. Specifically, this study addresses two key questions: (1) How does newcomers' collaboration network centrality in the new team influence their exploratory innovation performance within that team? (2) How does newcomers' knowledge network centrality in the new team moderate the above mechanism? We utilize global patent data from the European Patent Office's PATSTAT to identify instances of talent mobility, and then link these to the COMPUSTAT database for institutional disambiguation, ultimately obtaining 65,438 mobility records of technical talents.

Drawing on these newcomers' patent applications—both in their original and new teams—and on longitudinal patent data of the new teams, we construct measures for newcomers' collaboration network centrality, knowledge network centrality, and the team's exploratory innovation. We then employ negative binomial regression to test the proposed hypotheses.

Our empirical findings show that newcomers' centrality in the team's collaboration network exhibits an inverted U-shaped relationship with the team's exploratory innovation: at moderate levels of collaboration network centrality, newcomers can better balance the efficiency of information exchange and the costs of coordination, thus maximizing exploratory innovation; yet when centrality is either too high or too low, communication barriers, cognitive redundancy, or knowledge silos may arise, which inhibit team innovation performance. Further analyses reveal that knowledge network centrality negatively moderates this inverted U-shaped effect—when newcomers occupy higher positions in the knowledge network, the inverted U-curve becomes flatter, suggesting that individuals with rich knowledge resources maintain relatively stable innovation performance regardless of their collaboration network positions. This finding indicates that the "knowledge dimension" serves as a buffer that reduces the impact of the "collaboration dimension," enabling individuals with high knowledge network centrality to achieve consistent innovation outcomes across different collaborative contexts, while those with low knowledge network centrality are more sensitive to their collaboration network positions.

This study makes several important contributions. Theoretically, it first extends our understanding of how talent mobility influences team innovation, responding to scholarly debates regarding how external knowledge acquisition and network centrality interact to shape exploratory innovation. Second, by incorporating both collaboration networks and knowledge networks into the analysis of newcomer integration, it demonstrates that different dimensions of network centrality not only independently affect innovation but also alter the shape of the curve through interaction effects. Specifically, our findings reveal that knowledge network centrality flattens the inverted U-shaped relationship between collaboration network centrality and exploratory innovation, thus enriching our awareness of the boundary conditions of curvilinear effects under multiple variables. Moreover, this study underscores the pivotal role of individual-level network centrality in shaping team-level innovation, providing new empirical evidence for the micro–macro linkage in network theory. Practically, this study offers actionable guidance for managers in designing precise talent recruitment and integration strategies: organizations should consider newcomers' dual centrality in collaboration and knowledge networks, avoiding scenarios in which they become overly concentrated at the core, which can lead to resource redundancy or collaboration overload, as well as preventing them from being relegated to the periphery, resulting in insufficient support. Additionally, our findings suggest that firms can benefit from promoting cross-departmental collaboration and encouraging newcomers to engage extensively in various knowledge domains, thereby helping them build stronger "adhesion" in knowledge networks with broader coverage of expertise. Such approaches can help maintain stable innovation performance across different levels of collaboration network

centrality and enable organizations to better leverage external talents for enhanced exploratory innovation and sustained competitive advantage.

Literature Review and Hypotheses

Talent Mobility and Teams' Exploratory Innovation

Talent mobility and its impact on teams' exploratory innovation have emerged as significant areas of research in recent years. Exploratory innovation, characterized by substantial performance improvements, cost reductions, or addressing unmet needs, often disrupts existing markets or creates new ones, distinguishing itself from incremental innovation (Bower & Christensen, 1996; Subramaniam & Youndt, 2005). Talent mobility, defined as the movement of individuals within and across organizations, facilitates the transfer of knowledge, skills, and experiences (Kogut & Zander, 1992). This process is particularly crucial in high-tech industries, where it helps bridge technological gaps and accelerates advancements (Cascio & Montealegre, 2016).

The literature consistently highlights talent mobility's role in knowledge dissemination, resource integration, and the development of innovation ecosystems (Jotabá et al., 2022). Mobile high-skilled professionals carry both tacit and explicit knowledge, providing new technological pathways and innovation inspiration to receiving organizations through learning and imitation effects (Kerr et al., 2016). Furthermore, cross-industry, cross-cultural, or interdisciplinary mobility enables the integration of diverse knowledge backgrounds and cognitive models, fostering "knowledge collision" effects (Acar, Tarakci, & Van Knippenberg, 2019).

Two core mechanisms—collaboration networks and knowledge networks—are instrumental in this process. Collaboration networks connect previously isolated innovation actors, offering teams diverse resources and technical support while enhancing their cross-disciplinary collaboration capabilities (Newman, 2001). Knowledge networks, on the other hand, accelerate knowledge flow and sharing, enabling teams to integrate diverse perspectives and foster exploratory innovations (Phelps, Heidl, & Wadhwa, 2012). The synergy between these networks not only mitigates uncertainties associated with talent mobility but also expands the boundaries of the innovation ecosystem (Eslami, Ebadi, & Schiffauerova, 2013; Deichmann et al., 2020).

To further illustrate the interplay between collaboration and knowledge networks in the context of talent mobility, we present Figure 1, which depicts four possible collaboration strategies for newly recruited talents.

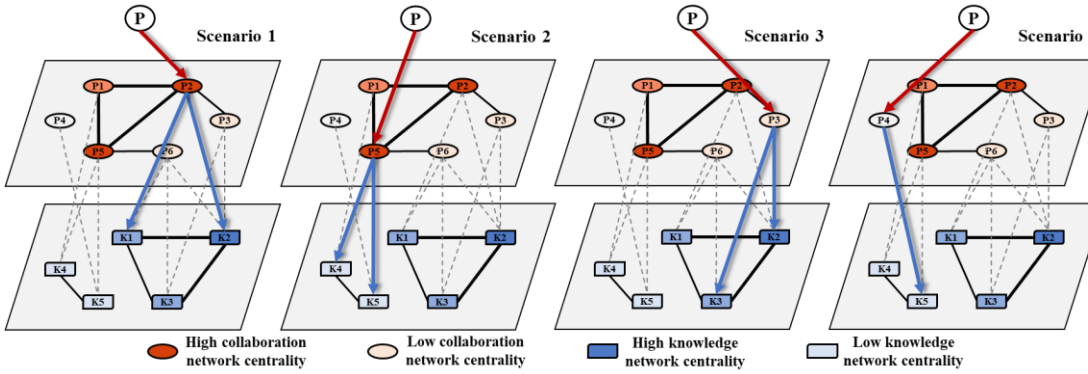


Figure 1. Schematic Diagram of Collaboration Strategies in Talent Mobility.

Figure 1 illustrates four scenarios that may arise when new talents join a team:

1) Newcomers occupy central positions in both the collaboration network and the knowledge network;
2) Newcomers are central in the collaboration network but peripheral in the knowledge network;
3) Newcomers are peripheral in the collaboration network but central in the knowledge network;
4) Newcomers occupy peripheral positions in both networks.

These scenarios highlight the complex relationship between collaboration network centrality and knowledge network centrality. While both types of centrality can contribute to innovation, their interaction may yield varied outcomes. For instance, when newcomers are central in both networks (scenario 1), they may be well-positioned to leverage their connections and expertise to drive exploratory innovation. However, this scenario might also lead to information redundancy or overload if not managed properly. Conversely, when newcomers are central in the collaboration network but peripheral in the knowledge network (scenario 2), they may facilitate information flow and resource allocation but might lack the specific expertise to substantially contribute to exploratory innovation. The opposite situation (scenario 3) could result in underutilized expertise if the newcomer's knowledge is not effectively integrated into the team's collaborative efforts.

In summary, talent mobility significantly influences exploratory innovation in teams through knowledge diffusion, team diversity enhancement, and resource reallocation. While previous research has extensively documented these effects, the specific role of network centrality in collaboration and knowledge networks during talent mobility has been underexplored. This study addresses this gap by focusing on how the centrality of newcomers in these networks impacts team exploratory innovation. The interplay between collaboration and knowledge network centrality serves as a critical mechanism in this process, facilitating knowledge transfer and organizational learning. By examining this relationship, we shed light on the complex dynamics underlying talent mobility and team innovation, offering new insights into how organizations can strategically leverage newcomers' network positions to enhance their innovative capabilities.

Collaboration Network Centrality and Teams' Exploratory Innovation

Collaboration networks, rooted in social network theory, have evolved into powerful analytical tools for understanding the structure and dynamics of scientific and organizational collaboration (Newman, 2001). These networks are characterized by nodes representing individuals or organizations, with edges signifying collaborative relationships such as co-authorship, joint projects, or advice-giving interactions (Camarinha & Afsarmanesh, 2005; Guimera et al., 2005). Key features of collaboration networks, including density, centrality, and connectivity, play crucial roles in influencing team innovation and performance (Van der Voet & Steijn, 2021). Recent research has highlighted the importance of examining the structural influence of centrality in these networks, particularly in the context of leadership and team effectiveness (Yuan & Van Knippenberg, 2022).

The relationship between collaboration network centrality and teams' exploratory innovation is complex and multifaceted, often contingent on various factors such as team size, organizational context, and the nature of the innovation tasks. Centrality, which measures a node's importance within a network, captures the extent to which an individual is connected to others and can influence information flow, resource access, and knowledge recombination (Tzabbar, Cirillo, & Breschi, 2022; Yang et al., 2021). In the context of newly recruited talents, their position in both collaboration and technological recombination networks can significantly impact their contribution to team innovation and their likelihood of remaining with the organization (Li et al., 2020).

This study proposes that the centrality of newly recruited talents within a team's collaboration network has a significant, inverted U-shaped effect on the team's exploratory innovation. This relationship can be explained through the interplay of two opposing mechanisms: knowledge integration and coordination costs. The knowledge integration mechanism positively influences exploratory innovation as centrality increases. As newly recruited talents become more central in the collaboration network, they gain greater access to diverse information, resources, and expertise within the team (Li et al., 2020; Bunderson, Van der Vegt, & Sparrowe, 2014). This enhanced access allows them to more effectively combine their unique perspectives with existing team knowledge, facilitating novel idea combinations and cross-pollination of concepts (McAdam & McClelland, 2002; Li, Mitchell, & Boyle, 2016). Conversely, the coordination costs mechanism negatively impacts exploratory innovation as centrality rises (Becker & Murphy, 1992). As newcomers become increasingly central, they face growing demands for coordination and communication with numerous team members (Srikanth & Puranam, 2014). This leads to potential information overload, increased cognitive strain, and the emergence of communication bottlenecks (Lingo, 2023). Higher centrality may lead to an imbalance in perceived power within the team. While the highly central newcomer might be more inclined to share knowledge due to their strong personal influence, other team members may experience a perceived loss of power. This can significantly reduce their willingness to share knowledge and potentially increase knowledge hiding behaviors, ultimately limiting the diversity of perspectives and ideas contributing to the innovation process (Issac et al., 2023).

The interplay of these two mechanisms creates the inverted U-shaped relationship. At low levels of centrality, the positive effects of knowledge integration are limited due to restricted access to team resources and information, while coordination costs are minimal. As centrality increases to moderate levels, the benefits of knowledge integration grow more rapidly than the coordination costs, creating an optimal balance where newcomers can effectively access and integrate diverse knowledge without being overwhelmed by excessive coordination demands. This balance maximizes their contribution to the team's exploratory innovation. However, when centrality increases beyond the optimal point, the negative effects of coordination costs begin to outweigh the positive effects of knowledge integration. The cognitive and communicative burdens of high centrality start to hinder the newcomer's ability to effectively process and utilize the wealth of information available, ultimately impeding the team's exploratory innovation performance. This inverted U-shaped relationship indicates that there is an optimal level of collaboration network centrality that maximizes exploratory innovation, where the positive effects of knowledge integration are maximized while the negative impacts of coordination costs are still manageable. Based on this, the hypotheses of this study are formulated as follows:

H1: *Newly recruited talents' collaboration network centrality exerts an inverted U-shaped effect on teams' exploratory innovation.*

Knowledge Network Centrality and Teams' Exploratory Innovation

Knowledge networks, distinct from yet interconnected with collaboration networks, play a crucial role in facilitating knowledge flow, integration, and innovation within organizations (Deichmann et al., 2020; Ren & Zhao, 2021). While collaboration networks emphasize interpersonal relationships, knowledge networks focus on the connections between knowledge elements and their dissemination processes (Phelps, Heidl, & Wadhwa, 2012). The centrality within knowledge networks reflects an individual's position in terms of access to and control over knowledge resources, which can significantly influence the dynamics of team innovation (Dong & Yang, 2016).

Building on the inverted U-shaped relationship established in the previous section, this study proposes that knowledge network centrality moderates the effect of collaboration network centrality on teams' exploratory innovation. The moderation effect can be explained by examining how knowledge network centrality influences the two underlying mechanisms - knowledge integration and coordination costs - across different levels of collaboration network centrality.

In the first phase of the inverted U-shaped relationship, where knowledge integration benefits dominate, high knowledge network centrality may attenuate the positive effect of increasing collaboration network centrality. Newly recruited talents with high knowledge network centrality already possess a rich knowledge base and extensive knowledge connections. Consequently, they may be less inclined to fully leverage the knowledge integration advantages offered by a central position in the collaboration network (Wang, Chen, & Fang, 2018). Instead, these individuals might

rely more heavily on their own expertise and knowledge resources to drive innovation (Lin et al., 2022). This self-reliance can lead to a reduced need for knowledge integration from team members, potentially diminishing the marginal utility of additional collaborative connections. Moreover, high knowledge network centrality may foster greater innovation autonomy, encouraging newcomers to pursue exploratory innovation independently rather than through extensive team collaboration (Guan & Liu, 2016; Wang & Yang, 2019).

In the second phase, where coordination costs become predominant, high knowledge network centrality may mitigate the negative effects associated with excessive collaboration network centrality. Newcomers with high knowledge network centrality are likely to possess deep domain expertise, enabling them to more efficiently process and integrate information from various team members (Dong & Yang, 2016; Guan, Yan, & Zhang, 2017). This expertise can lead to more effective communication, as these individuals can quickly identify and focus on critical information, reducing unnecessary coordination efforts (Jiang, Shi, & Cheng, 2024). Furthermore, their extensive knowledge base may allow them to solve problems more independently, decreasing their reliance on other team members and thus lowering overall coordination demands (Tang, Fang, & Qualls, 2020). High knowledge network centrality may also enable newcomers to focus their innovation efforts within their areas of expertise, potentially reducing the need for cross-domain coordination and its associated costs (Wang & Zheng, 2022).

The combined effect of these moderation processes on both phases of the inverted U-shaped relationship is a flattening of the overall curve. This flattening suggests that individuals with high knowledge network centrality maintain relatively stable innovation performance across different levels of collaboration centrality. Their extensive knowledge resources and integration capabilities allow them to contribute effectively to exploratory innovation even when their collaboration network centrality is suboptimal (Guan & Liu, 2016; Wang et al., 2014). Based on this analysis, we formulate the following hypothesis:

H2: *Knowledge network centrality moderates the inverted U-shaped relationship between newly recruited talents' collaboration network centrality and teams' exploratory innovation, such that higher knowledge network centrality attenuates this curvilinear relationship—making the inverted U-shaped curve flatter.*

Overall of the Conceptual Framework

Figure 2 presents our research model, focusing on newly recruited talents and their impact on team exploratory innovation. The model illustrates the interplay between collaboration network centrality, knowledge network centrality, and innovation outcomes.

In our research context, newly recruited talents enter teams with varying degrees of centrality in both collaboration and knowledge networks. The collaboration network centrality of these newcomers has an inverted U-shaped effect on team exploratory innovation, driven by the balance between knowledge integration benefits and coordination costs. As collaboration centrality increases from low to moderate levels,

knowledge integration benefits dominate, enhancing innovation. However, beyond an optimal point, coordination costs become more pronounced, leading to a decline in innovation outcomes. The knowledge network centrality of newly recruited talents moderates this inverted U-shaped relationship, attenuating its curvature. High knowledge network centrality flattens the relationship by dampening both the positive effects of knowledge integration and the negative effects of coordination costs. This suggests that individuals with high knowledge network centrality maintain relatively stable innovation performance across different levels of collaboration centrality. Our dual-network perspective integrates collaboration and knowledge dimensions, offering a comprehensive view of how talent mobility and network positions influence team innovation.

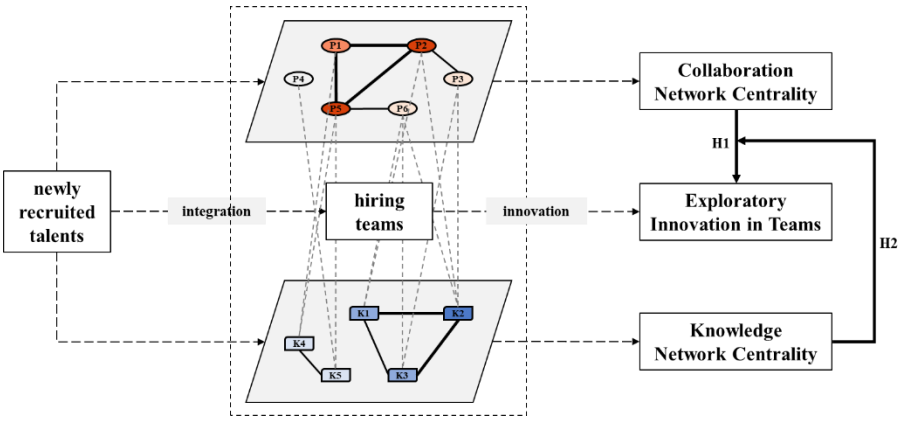


Figure 2. Research Model.

Data and Methods

Sample Selection

This study utilizes data from the European Patent Office's (EPO's) PATSTAT (2020 Spring edition), a comprehensive global patent database widely employed in innovation and patent analysis research (Wang et al., 2024; Shi et al., 2023). PATSTAT provides extensive bibliographic information on patent applications and publications worldwide since 1978. To identify talent mobility events, we track the movement of technical personnel by examining consecutive patent application records where the assignee changes, indicating a shift from one organization to another (Singh & Agrawal, 2011). Specifically, an inventor is considered to have moved when there is a change in the assignee between two successive patent applications. The midpoint between the filing dates of these two patents is used as an estimated mobility time (Song et al., 2003).

To address data ambiguities and redundancies, such as firm renaming or restructuring, we cross-reference PATSTAT data with the COMPUSTAT database, which provides detailed information on companies traded on U.S. or Canadian exchanges. Following the methodology established in prior studies (Bessen, 2008), we disambiguate firm names by matching identification fields between PATSTAT

and COMPUSTAT, successfully resolving ambiguities caused by name changes, mergers, acquisitions, or parent-subsidiary relationships.

After the disambiguation process, we applied several filters to ensure the reliability and relevance of our sample. We focused on mobility events where each inventor moved only once, avoiding complications related to short observation windows and insufficient innovation data. To guarantee established collaboration networks and innovation foundations, we required inventors to have at least two patent applications in both their original and new teams. We restricted the time gap between consecutive patent applications to 2-5 years, allowing for accurate estimation of mobility timing while excluding events with potentially inaccurate identification due to short time gaps. Our study concentrated on mobility events occurring between 1996 and 2010, providing a sufficient window to observe subsequent knowledge transfer and innovation outcomes.

In addition to these primary filters, we exclude outliers to enhance data quality: inventors with an unusually large number of patent applications, those with exceptionally long technological careers (e.g., over 90 years), and those who receive an abnormally high number of citations before and after moving. These exclusions help mitigate the effects of atypical cases that could distort the analysis. After applying these stringent criteria, the final sample consists of 65,438 mobility events. This refined sample ensures that the impact of talent mobility on exploratory innovation can be accurately assessed within teams that have a pre-existing collaboration network and innovation capacity, thereby enhancing the validity and reliability of our empirical findings.

Dependent Variable

Teams' exploratory innovation measures the extent to which teams develop novel knowledge and technologies that significantly enhance performance, reduce costs, or address unmet needs. To accurately capture exploratory innovation, this study utilizes patent data co-applied by newly recruited technical personnel and their collaborators within the team.

Exploratory innovation is operationalized by analyzing patents filed within five years following a talent mobility event (t+1 to t+5 years). These patents are compared against those filed in the five years preceding the mobility event (t-1 to t-5 years) using the International Patent Classification (IPC) codes, which represent the knowledge elements within the team. A patent filed in the post-mobility period is classified as an exploratory innovation if it includes IPC codes not present in the pre-mobility period. The total frequency of these new IPC codes serves as the measure of exploratory innovation, with a higher frequency indicating a greater extent of innovative activities introduced by the newly recruited talents. To ensure the reliability and relevance of the measurements, only patents directly co-applied by the moving technical personnel and their immediate collaborators are included, ensuring that the patents reflect the direct contributions of the newly recruited talents to the team's innovation efforts.

Independent Variable

The primary independent variable in this study is collaboration network centrality (Cnc), which quantifies the position of newly recruited talents within the team's collaboration network. Cnc measures the extent to which a talent is embedded within influential and interconnected segments of the collaboration network, reflecting their ability to facilitate effective knowledge transfer and foster innovative collaborations. Specifically, Cnc is assessed by calculating the mean eigenvector centrality of all collaborators associated with the newly recruited talent over the five-year period preceding their mobility event (from t−5 to t). Eigenvector centrality is chosen for its capacity to capture not only the number of direct connections a collaborator has but also the quality and influence of those connections within the network (Dong & Yang, 2016). By averaging the eigenvector centrality scores of all collaborators, Cnc provides a comprehensive measure of a talent's overall influence and integration within the collaboration network, thereby serving as a robust indicator of their potential to drive exploratory innovation within the team. The mean eigenvector centrality for Cnc is calculated as follows:

$$Cnc_i = \frac{1}{N_i} \sum_{j=1}^{N_i} C_{ij}^{eigenvector} \quad (1)$$

$\mathrm{Cnc_i}$ is the collaboration network centrality of the i-th newly recruited talent. $\mathrm{N_i}$ is the number of direct collaborators of the i-th talent within the collaboration network. $C_{ij}^{\mathrm{eigenvector}}$ represents the eigenvector centrality of the j-th collaborator connected to the i-th talent.

Moderator Variable

The moderator variable in this study is knowledge network centrality (Knc), which measures the position of newly recruited talents within the team's knowledge network. Similar to Cnc, Knc assesses the influence and integration of a talent within the knowledge flow processes of the team. Knc is determined by calculating the mean eigenvector centrality of all collaborators associated with the newly recruited talent in the knowledge network over the same five-year period (from t-5 to t). This measure captures the extent to which a talent is embedded within a highly influential knowledge network, facilitating efficient knowledge dissemination and integration. By averaging the eigenvector centrality scores of all knowledge collaborators, Knc serves as an indicator of the talent's ability to enhance the team's innovation capacity through effective knowledge management and integration. The mean eigenvector centrality for Knc is calculated as follows:

$$Knc_i = \frac{1}{M_i} \sum_{k=1}^{M_i} K_{ik}^{eigenvector} \quad (2)$$

Knc_i is the knowledge network centrality of the i-th newly recruited talent. M_i is the number of direct knowledge collaborators of the i-th talent within the knowledge network. $K_{ik}^{\mathrm{eigenvector}}$ represents the eigenvector centrality of the k-th knowledge collaborator connected to the i-th talent.

Control Variables

To ensure that the effects of Cnc and Knc on teams' exploratory innovation are not confounded by other factors, this study incorporates several control variables categorized into three dimensions: characteristics of newly recruited talents, characteristics of new teams, and relational dynamics between talents and teams.

In terms of the newly recruited talents' characteristics, the study first measures the work experience of the talent. This is calculated as the number of years between the earliest patent application year of the talent and the year of their mobility event (Talent Age, Ta). A longer work age indicates greater experience, potentially enhancing the talent's ability to contribute to team innovation. Additionally, the study considers the total number of patents the talent has applied for prior to their mobility event (Talent Patent Number, Tpn). This variable serves as an indicator of the talent's accumulated technical innovation experience. The research also examines the average number of collaborators the talent has worked with on past patents before moving (Talent Social Capital Average, Tsc). This metric reflects the talent's ability to engage in collaborative innovation and leverage social networks within the team. Furthermore, the study assesses the average position of the talent in their past collaborative patents (Talent Knowledge Capital Average, Tkc). A higher average position indicates greater knowledge importance and capital, signifying the talent's influential role in collaborative endeavors.

Regarding the characteristics of new teams, the study includes a count of the number of patents the new team has filed in the five years preceding the talent mobility event (New Team Patents Base In5, Ntpb). This measures the team's existing knowledge base and innovation capacity prior to the influx of new technical personnel. Additionally, the total number of technical personnel in the new team before the mobility event is considered (New Team Talent Number, Nttn). This controls for team size and the team's experience in managing collaborations and innovation processes.

In terms of the relational dynamics between talent and team, the study incorporates a binary variable indicating whether the new team has previously cited the talent's patents in their own patents before the mobility event (Prior Cites, Pc). This captures pre-existing knowledge links that may influence collaborative strategies post-mobility. Additionally, the total number of collaborators the newly recruited talent has in the new team after the mobility event is counted (Co-inventor Count, Cic). This controls for the extent of collaborative interactions, which can directly influence the team's innovative activities. All variables and their description are shown in Table 1.

Table 1. Variables Description.

Variable	*Abbreviation*	*Description*
Dependent		
Exploratory Innovation in Teams	Exploratory	Measured by the total number of new IPC codes introduced in patents filed by the team within five years following a talent mobility event. A higher count indicates a greater extent of exploratory innovation driven by newly recruited talents.
Independent		
Collaboration Network Centrality	Cnc	Calculated as the average eigenvector centrality of all collaborators associated with the newly recruited talent over the five years preceding their mobility event. This metric reflects the talent's overall influence and integration within the collaboration network.
Moderator		
Knowledge Network Centrality	Knc	Determined by the average eigenvector centrality of all knowledge collaborators connected to the newly recruited talent over the same five-year period. It indicates the talent's position and influence within the knowledge network, facilitating effective knowledge flow and integration.
Control		
Talent Age	Ta	Calculated as the number of years between the earliest patent application year of the talent and the year of their mobility event. This measures the work experience of the newly recruited talent.
Talent Patent Number	Tpn	Represents the total number of patents the talent has applied for prior to their mobility event, indicating their accumulated technical innovation experience and expertise.
Talent Social Capital Average	Tsc	Measures the average number of collaborators the talent has worked with on past patents before moving, reflecting their ability to engage in collaborative innovation and leverage social networks within the team.
Talent Knowledge Capital Average	Tkc	Assesses the average position of the talent in their past collaborative patents. A higher average position signifies greater knowledge importance and capital, indicating the talent's influential role in collaborative endeavors.
New Team Patents Base In5	Ntpb	Counts the number of patents the new team has filed in the five years preceding the talent mobility event, serving as a measure of the team's existing knowledge base and innovation capacity prior to the influx of new technical personnel.
New Team Talent Number	Nttn	Represents the total number of technical personnel in the new team before the mobility event, controlling for team size and the team's experience in managing collaborations and innovation processes.
Prior Cites	Pc	A binary variable indicating whether the new team has previously cited the talent's patents in their own patents before the mobility event. It captures pre-existing knowledge links that may influence collaborative strategies post-mobility.
Co-inventor Count	Cic	Counts the total number of collaborators the newly recruited talent has in the new team after the mobility event, controlling for the extent of collaborative interactions that can directly influence the team's innovative activities.

Results

Descriptive Statistical Analysis of Variables

Before conducting the empirical analyses, we first present the descriptive statistics of all key variables employed in this study. This section provides the means, standard deviations (SD), correlation coefficients, and variance inflation factors (VIF) for both the focal and control variables. Table 2 contains a detailed overview of these statistics.

As shown in Table 2, the mean value of Exploratory is 81.41, with a standard deviation of 346.52. This relatively large standard deviation indicates substantial variation among teams in terms of their exploratory innovation outputs—some teams demonstrate markedly higher innovation performance due to greater resource inputs or stronger R&D capabilities, whereas others may be more constrained in these areas. Given the nature of our dependent variable, we employ a negative binomial regression model for empirical testing. This choice is justified by the characteristics of the Exploratory variable, which exhibits overdispersion. Specifically, the variance is significantly larger than the mean, indicating that a Poisson regression would not be suitable for effective empirical analysis. The negative binomial model is better equipped to handle this overdispersion, providing more accurate estimates and reducing the risk of biased standard errors that could lead to incorrect inferences about the significance of our predictors.

Regarding the key independent variables, the mean of Cnc is 0.29 (SD = 0.41), and the mean of Knc is 0.36 (SD = 0.37). These statistics suggest that newly hired talents vary considerably in how centrally they are positioned in the team's collaboration and knowledge networks—some newcomers quickly occupy more central roles, while others remain on the periphery. Examining the correlations, several notable findings align with our theoretical expectations. First, Cnc is positively and significantly correlated with Exploratory ($r = 0.10$, $p < 0.001$), indicating that a more central position in the collaboration network tends to be associated with higher levels of exploratory innovation. Cnc is also moderately and significantly correlated with Knc ($r = 0.37$, $p < 0.001$), suggesting that newcomers who occupy prominent positions in the collaboration network often hold similarly central positions in the knowledge network.

Finally, the variance inflation factor (VIF) values are generally low (all below 3), with the highest being 2.62 for Ta, well under typical cutoffs (5 or 10). Hence, multicollinearity is unlikely to pose a serious issue in our regressions. Overall, these descriptive statistics and correlations lend preliminary support to our hypotheses regarding the importance of newcomers' network positions for achieving higher levels of exploratory innovation, and they set the stage for the subsequent regression analyses.

Table 2. Correlation Analysis.

Variables	*Mean*	*SD*	*VIF*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Exploratory	81.41	346.52	/	1.00										
Cnc	0.29	0.41	1.25	0.10***	1.00									
Knc	0.36	0.37	1.20	0.00	0.37***	1.00								
Pc	0.13	0.34	1.12	0.02***	0.03***	0.06***	1.00							
Cic	2.83	3.53	1.09	0.35***	0.12***	0.13***	0.11***	1.00						
Ta	4.23	4.28	2.62	-0.06***	-0.10***	-0.06***	-0.10***	-0.12***	1.00					
Tpn	5.61	8.38	2.45	0.02***	0.01***	0.00	0.11***	0.04***	0.25***	1.00				
Tsc	3.60	2.58	1.48	0.11***	0.11***	0.15***	0.04***	0.37***	-0.08***	0.05**	1.00			
Tkc	2.27	1.62	1.60	0.09***	0.09***	0.11***	0.02***	0.29***	-0.10***	0.01**	0.77**	1.00		
Ntpb	45.40	93.46	1.04	-0.01**	-0.22***	-0.14***	-0.04***	0.02***	0.09***	0.01*	0.02**	0.02**	1.00	
Nttn	29.95	49.35	1.25	0.05***	-0.24***	-0.14***	0.03***	0.22***	-0.01*	0.02**	0.13**	0.10**	0.55**	1.00

* p < 0.05; ** p < 0.01; *** p < 0.001

Data Distribution Analysis

Figure 3 presents the overall distribution of the data in this study, illustrated through two subplots: (a) the number of talent mobility events within three-year intervals and (b) the average level of exploratory innovation (i.e., patent-based metric) per year. These figures help to contextualize the temporal trends in talent mobility and subsequent innovation outcomes, as well as provide preliminary insights into how broader external factors might have shaped these patterns over time.

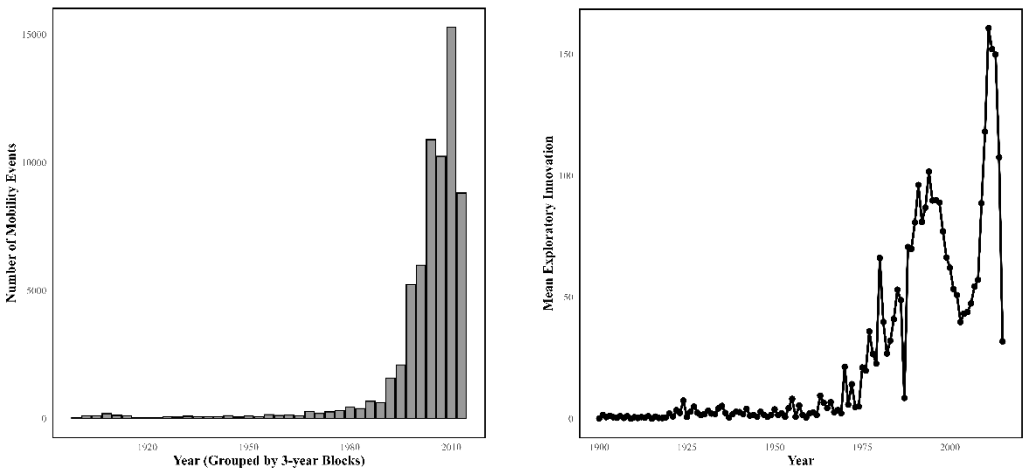
Figure 3a displays the frequency of talent mobility events using three-year windows. The data indicate a steady rise in mobility around the early 1990s, accelerating more sharply between 2000 and 2005, followed by a peak around 2010. Several possible factors could have driven this trajectory. During the late 1990s and early 2000s, the dot-com boom and the broader emergence of high-technology industries likely fueled the demand for specialized R&D talent. As startups proliferated and established firms invested aggressively in innovation, mobility events naturally increased. The 1990s and early 2000s saw rapid globalization, with multinational corporations expanding their operations worldwide. This environment created numerous international collaborations and cross-border R&D teams, which in turn heightened the movement of technical professionals. Around the 2010 peak, firms were recovering from the financial downturn of 2008–2009 and making strategic investments in new research fields. As companies reorganized and diversified, the recruitment of external talent became a focal strategy, pushing mobility events to a high point.

Figure 3b tracks the variation in teams' average exploratory innovation outputs across different time periods, capturing how new hires contributed to cutting-edge R&D. The figure reveals several notable fluctuations. Around 1975, there is an initial surge in exploratory patenting activities. One possible explanation is the heightened innovation impetus driven by government support and industrial restructuring post–World War II, which continued to foster both technology advancement and talent mobility. Another significant uptick occurs around 1995, potentially corresponding to the mainstream adoption of personal computing, the internet's early commercial phase, and broader transitions in telecommunications technology. Together, these

trends likely spurred new patenting opportunities and incentivized firms to acquire external talent with specialized expertise. Following the 1995 spike, a noticeable dip appears around 2000. This decline may reflect the burst of the dot-com bubble, which led to reduced R&D spending in certain sectors and a slowdown in venture funding. Consequently, the intensity of exploratory patenting could have temporarily contracted during this period of market readjustment. Subsequently, exploratory innovation spikes again from 2005 to 2010, possibly reflecting the advent of new technologies and a revitalized venture capital environment. Many firms resumed or intensified R&D investments, actively recruiting technical talents from various fields to strengthen their innovative capabilities.

Taken together, these patterns underscore both the cyclical nature of technology-driven industries and the strong link between external shocks and fluctuations in talent mobility and subsequent innovation activities. The significant peaks in mobility and exploratory innovation suggest that firms not only capitalized on buoyant markets to expand their human capital but also recognized the strategic importance of injecting novel knowledge into their existing R&D processes. Conversely, during economic downturns or after market corrections, fewer mobility events and reduced innovation outputs may indicate contraction in research investment or a more cautious approach to integrating new technological avenues.

These descriptive insights reinforce the importance of examining how newly hired talents' network positions can help—or hinder—teams in realizing exploratory innovation. As the broader historical context implies, successful talent mobility appears to depend upon both external environmental factors and the newcomers' ability to leverage their social and knowledge connections once integrated into the team.



(a) Mobility Events in 3-year Intervals (b) Average Exploratory Innovation per Year

Figure 3. Data Distribution.

Visualization of the Mobility Network

Figure 4 employs a network visualization to depict the overall pattern of talent movement across different organizations in our sample. Here, each node represents

an organization, and each edge shows the aggregated number of individuals who transferred between two organizations. The resulting network is relatively sparse and dispersed, with only a few organizations standing out as central nodes—those that either attract or dispatch larger numbers of talent. These core nodes form only a handful of sub-networks, whereas most organizations remain separate from these clusters.

This dispersion suggests that talent mobility in our sample is not dominated by any single group of firms; rather, it is spread across a wide range of organizations, each with relatively distinct and independent flows of human capital. As a result, our data collection captures a more general mobility context rather than focusing on a narrow set of interconnected players. The relative sparsity of the network also provides reassurance regarding the randomness and representativeness of our sample, given that it does not overly concentrate on a small set of high-traffic channels.

Moreover, this visualization sheds light on the nature of international talent flows, revealing that even though some organizations serve as prominent "hubs," the broader pattern is one of dispersed and heterogeneous connectivity. This fragmentation reinforces the importance of understanding how new hires integrate and leverage their social and knowledge networks once they transition to a new team. Policymakers and managers interested in strengthening talent pipelines and innovation networks can draw on such insights to better design recruitment and collaboration strategies, recognizing that large-scale talent clusters are only one component of a more complex and widely distributed mobility landscape.

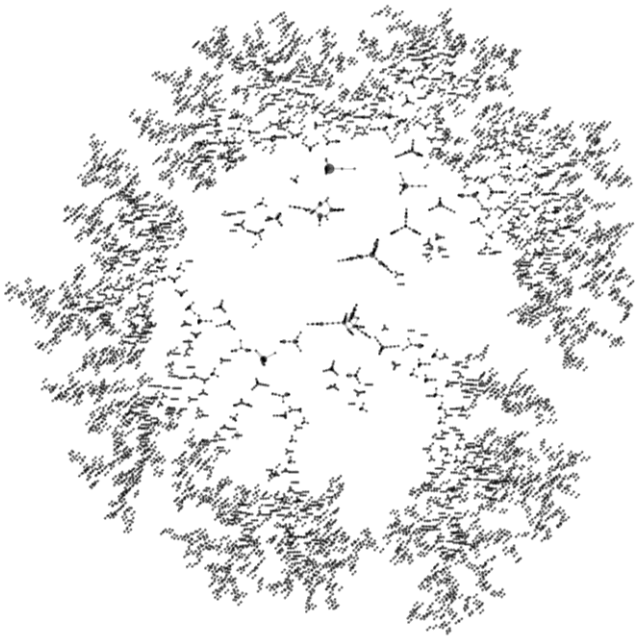


Figure 4. Mobility Network.

Network Position Correlation Relationship Analysis

Figure 5 provides a scatterplot of Cnc on the horizontal axis and Knc on the vertical axis, offering a visual representation of how these two variables co-vary across the mobility events in our dataset. Several observations stand out.

A substantial proportion of sample points fall in the upper-right quadrant of the scatterplot, suggesting that many newly recruited talents achieve both high collaboration network centrality and high knowledge network centrality within their new teams. This pattern is consistent with individuals who not only maintain active and diverse social ties but also command extensive or specialized knowledge resources. Despite the concentration in the high–high quadrant, there remain numerous cases in which newly hired talents exhibit a high level of Knc alongside a relatively low Cnc (upper-left quadrant) or a high Cnc with a relatively low Knc (lower-right quadrant). Additionally, some observations appear in the lower-left quadrant, characterized by both low collaboration and low knowledge centrality. These distributions validate our earlier conceptual typology in Figure 1, which proposed four distinct modes of newcomer integration based on the intersection of their positions in collaboration and knowledge networks. To further interpret these patterns, we draw on the mean values of Cnc and Knc to demarcate four quadrants, each reflecting a unique combination of collaboration and knowledge network positions. Assigning all sample points into these four categories helps illustrate that the hypothesized patterns of newcomer integration indeed emerge in practice and are not merely theoretical constructs.

Collectively, the distribution in Figure 5 underscores the heterogeneity of network positions occupied by newly recruited talents. While many newcomers manage to establish both broad social ties and access to rich knowledge resources, some may focus more on integrating into the knowledge structure before cultivating widespread collaboration links. This diversity of integration pathways reinforces the notion that talent mobility outcomes are shaped by a dynamic interplay between how individuals form social connections and how they leverage or contribute specialized knowledge. As our subsequent analyses will reveal, such differences in network positions can have significant implications for the level and nature of exploratory innovation within teams.

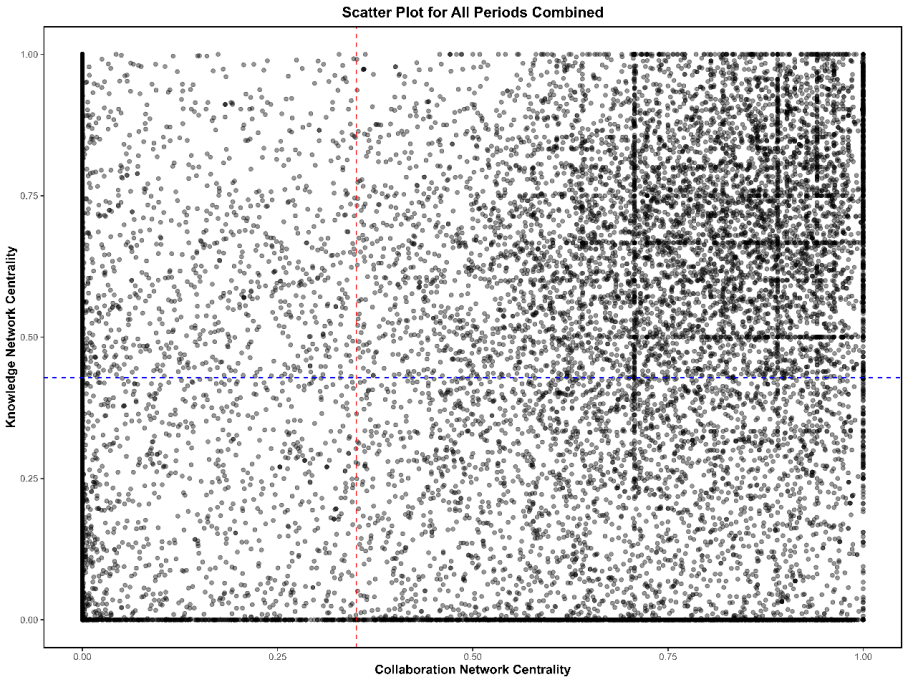
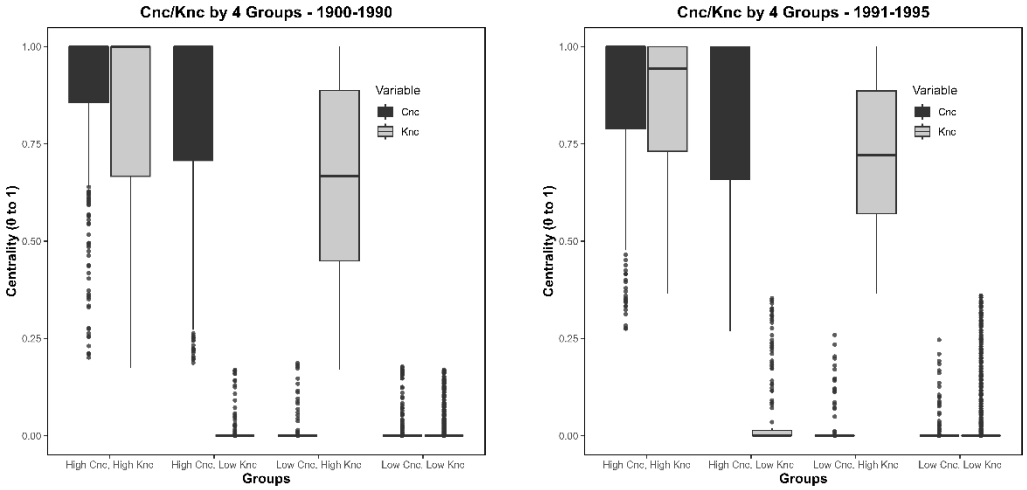


Figure 5. Scatterplot of Cnc vs. Knc.



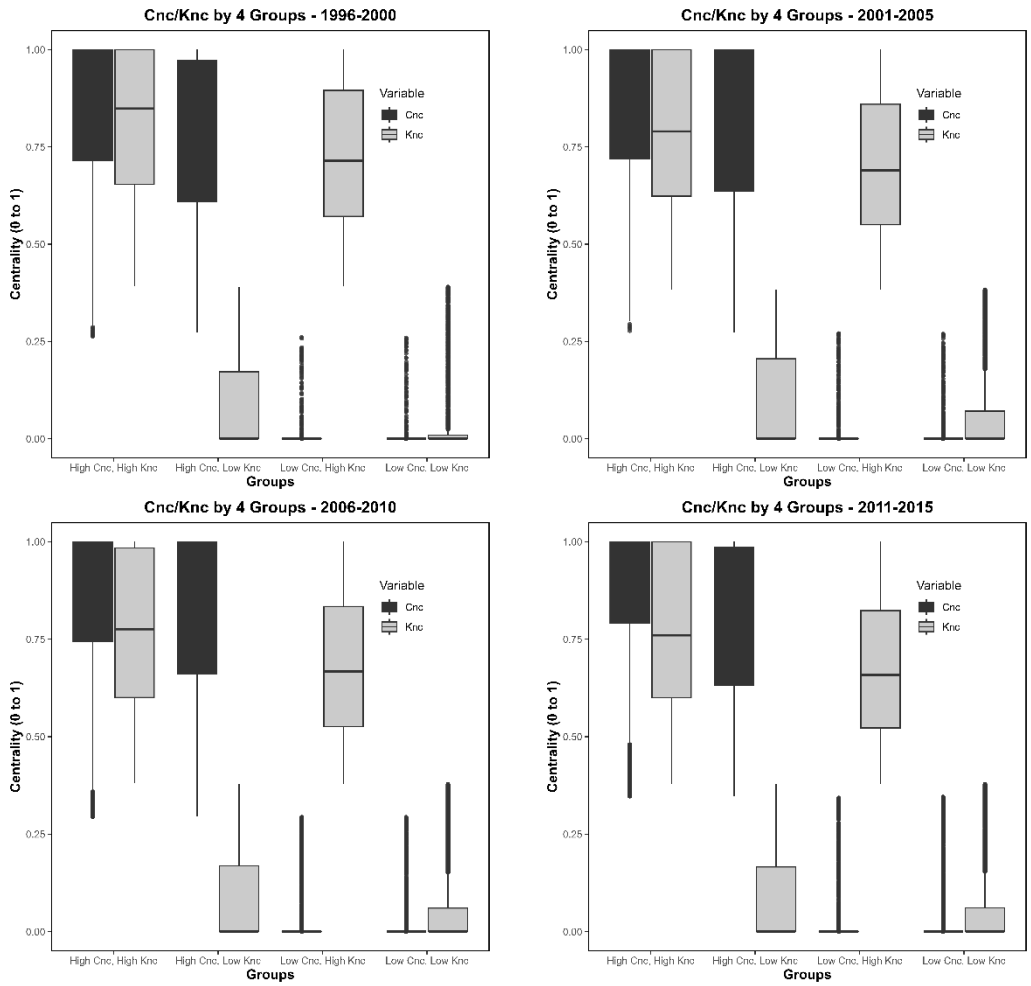


Figure 6. Boxplot of Cnc/Knc by 4 Groups.

Figure 6 extends the four-quadrant classification of newcomer integration by illustrating how these distinct categories—high–high, high–low, low–high, and low–low, in terms of Cnc and Knc—shift over time. For each of the six specified periods (1900–1990, 1991–1995, 1996–2000, 2001–2005, 2006–2010, and 2011–2015), boxplots reveal both the median and overall range of Cnc and Knc distributions within each group.

In the group of newcomers who occupy high Cnc and high Knc positions, the initial data indicate that individuals in this category consistently exhibit robust levels of both collaborative and knowledge-based embeddedness. As time progresses, however, there is a visible downward movement in the centers of both distributions, suggesting that the intensity of "double-core" embeddedness may have declined, possibly in response to more distributed organizational structures or a broader dispersion of expertise. Interestingly, in the 2011–2015 window, the central tendencies of this quadrant rebound slightly, hinting that recent waves of technology development or shifting organizational strategies may once again favor newcomers who achieve both high collaboration and high knowledge positions.

A contrasting picture emerges for those with high Cnc but low Knc. Although these newcomers initially register relatively modest knowledge centrality, the data show a gradual upward shift in their Knc values over successive periods. This movement suggests that individuals who are adept at building social connections within a team may subsequently gain or develop technical expertise, whether through training, mentoring, or project-based learning. By contrast, those with low Cnc but high Knc remain on the periphery of social collaborations throughout most timeframes, despite consistently holding a relatively strong knowledge base. Although they are not as embedded in collaboration networks as the high–high group, they still possess more specialized expertise than the low–low quadrant, pointing to a narrower, perhaps more specialized integration strategy in the team context.

The final quadrant, composed of individuals with both low Cnc and low Knc, registers a more limited capacity for either social engagement or technical contribution in the early periods of the sample. Yet after 2000, a noticeable increase appears in their median Knc values, suggesting that at least part of this group may be acquiring greater technical know-how over time. This shift could reflect a changing innovation climate, where even newcomers who start off with limited collaboration ties and knowledge resources can improve their standing if organizations provide relevant training or assign them to projects that facilitate skill development.

Taken together, these temporal boxplot patterns highlight the dynamic nature of newcomers' positions in both collaboration and knowledge networks. While some individuals maintain persistently high or low positions, the data also reveal that many evolve over time, reflecting shifts in industry priorities, organizational structures, and personal career trajectories. Understanding these trends is therefore essential for clarifying how talent mobility contributes to team-level innovation capacity, as high Cnc and Knc may be prized more strongly during certain technology cycles, whereas in other periods, the gradual elevation of knowledge among socially well-connected newcomers might become the dominant driver of exploratory R&D outcomes.

Empirical Estimation

Table 3 presents the results of the negative binomial regression models used to predict Exploratory. Model (1) includes only control variables. In Model (2), we add the key independent variable Cnc and its squared term (Cnc^2) to test the hypothesized inverted U-shaped relationship. Finally, in Model (3), we incorporate the moderating variable Knc and its interaction effects with both Cnc and Cnc^2.

The results in Model (2) provide clear evidence of an inverted U-shaped main effect. The coefficient for Cnc is 2.40 ($p < 0.01$), indicating that, up to a certain point, higher collaboration centrality is associated with greater team-level exploratory innovation. However, the coefficient for Cnc^2 is -1.60 ($p < 0.01$), suggesting that once Cnc surpasses a moderate level, its positive effect on exploratory innovation diminishes and eventually turns negative. This finding is in line with our theoretical argument that newcomers who are too central in the collaboration network may encounter communication overload or redundancy, whereas those who are too peripheral lack sufficient information exchange to drive breakthrough ideas.

To further validate the inverted U-shaped relationship, we calculated the inflection point of the curve. This inflection point (0.75) falls within the variable range of Cnc [0,1], confirming that the inverted U-shaped relationship is indeed observable within the scope of our data. The positive effect of collaboration network centrality on exploratory innovation reaches its peak when Cnc is at 0.75, after which the effect begins to decline. This finding provides strong support for our hypothesis and underscores the importance of achieving an optimal level of collaboration centrality to maximize team-level exploratory innovation. Hypothesis H1 is confirmed.

Table 3. Regression Results.

Variables	*Dependent variable: Exploratory*		
	Model 1	*Model 2*	*Model 3*
Cnc		2.40*** (0.10)	4.20*** (0.16)
Cnc^2		-1.60*** (0.11)	-3.10*** (0.18)
Knc			0.33*** (0.03)
Cnc×Knc			-5.00*** (0.29)
Cnc^2×Knc			4.50*** (0.30)
Pc	-0.13*** (0.02)	-0.12*** (0.02)	-0.11*** (0.02)
Cic	0.41*** (0.002)	0.38*** (0.002)	0.38*** (0.002)
Ta	-0.03*** (0.002)	-0.01*** (0.002)	-0.01*** (0.002)
Tpn	0.02*** (0.001)	0.01*** (0.001)	0.01*** (0.001)
Tsc	0.02*** (0.005)	0.02*** (0.005)	0.02*** (0.005)
Tkc	0.02*** (0.01)	0.01 (0.01)	0.01 (0.01)
Ntpb	0.0004*** (0.0001)	0.001*** (0.0001)	0.001*** (0.0001)
Nttn	-0.001*** (0.0002)	0.002*** (0.0002)	0.002*** (0.0002)
Constant	2.50*** (0.02)	2.10*** (0.02)	2.00*** (0.02)

The standard errors are shown in brackets, the same as below.
*, **, *** respectively represent $p < 0.1$, $p < 0.05$, $p < 0.01$

Model (3) tests the moderation effects by adding Knc and its interaction terms with Cnc and Cnc^2. Knc on its own has a significant positive effect on Exploratory (coefficient = 0.33, $p < 0.01$), illustrating that newcomers with broader or deeper knowledge connections can enhance a team's capacity for innovative outputs. More importantly, the interaction between Cnc and Knc is significantly negative (coefficient = -5.00, $p < 0.01$), and the interaction between Cnc^2 and Knc is significantly positive (coefficient = 4.50, $p < 0.01$).

These results provide support for our hypothesis regarding the moderating effect of knowledge network centrality. The positive interaction between Cnc^2 and Knc indicates that higher knowledge network centrality mitigates the negative quadratic effect of collaboration network centrality. In practical terms, these findings imply that the inverted U-shaped relationship between collaboration network centrality and exploratory innovation becomes flatter as knowledge network centrality increases. This means that for newcomers with high knowledge network centrality, the benefits of moderate collaboration network centrality are less pronounced, but the negative effects at extreme levels of collaboration centrality are also less severe.

To further illustrate this moderating effect, we have plotted the interaction in Figure 7. As shown in the figure, the inverted U-shaped relationship between collaboration network centrality and exploratory innovation becomes noticeably flatter when Knc is higher. This visual representation clearly demonstrates that as newcomers' knowledge network centrality increases, the curvilinear effect of their collaboration network centrality on team-level exploratory innovation becomes less pronounced. The graph underscores our finding that a high level of knowledge network centrality can buffer against the potential negative effects of both very low and very high collaboration network centrality, leading to a more stable relationship between collaboration centrality and exploratory innovation across different levels of Cnc.

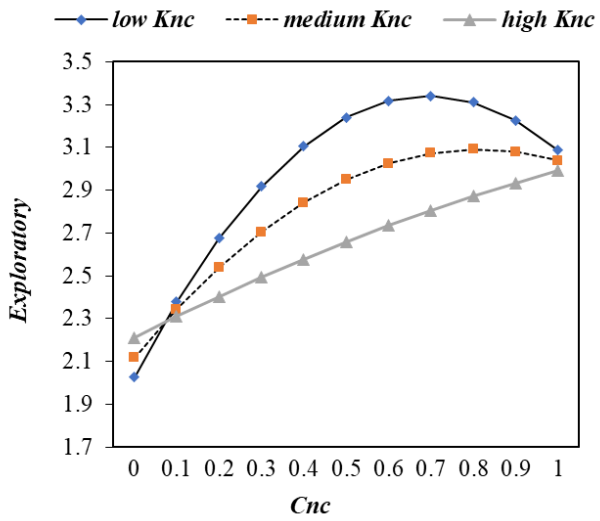


Figure 7. Moderating Effect Diagram.

Results

This study provides valuable insights into the complex dynamics of talent mobility, network centrality, and team-level exploratory innovation. Our findings contribute to both theoretical understanding and practical implications in several key areas.

Firstly, our results confirm the inverted U-shaped relationship between newcomers' collaboration network centrality and teams' exploratory innovation. This finding extends the existing literature on social networks and innovation (Newman, 2001; Li et al., 2020) by demonstrating that the benefits of network centrality are not linear but rather have an optimal point. At moderate levels of centrality, newcomers can effectively integrate diverse knowledge and resources, fostering innovation. However, excessive centrality can lead to coordination costs that outweigh these benefits, aligning with previous research on the cognitive limits of collaboration (Srikanth & Puranam, 2014; Lingo, 2023). This nuanced understanding of the centrality-innovation relationship has important implications for team composition and management in innovative organizations. It suggests that managers should strive for a balanced approach when integrating new talents, ensuring they have sufficient

connections to access diverse knowledge without becoming overburdened by excessive coordination demands.

Secondly, our study reveals the significant moderating role of knowledge network centrality on the relationship between collaboration network centrality and exploratory innovation. This finding contributes to the growing body of research on the interplay between different types of networks in organizational settings (Deichmann et al., 2020; Ren & Zhao, 2021). By demonstrating that high knowledge network centrality flattens the inverted U-shaped relationship, we highlight the importance of considering both collaboration and knowledge dimensions when studying innovation dynamics. This moderation effect suggests that individuals with high knowledge network centrality can maintain relatively stable innovation performance across different levels of collaboration centrality. This finding has practical implications for talent acquisition and team formation strategies. Organizations might benefit from prioritizing individuals with high knowledge network centrality, as they appear more resilient to suboptimal positioning within collaboration networks.

Our research also contributes to the broader discussion on talent mobility and innovation ecosystems (Jotabá et al., 2022; Cascio & Montealegre, 2016). By focusing on the network positions of newly recruited talents, we provide a more nuanced understanding of how organizations can leverage talent mobility to enhance their innovative capabilities. This perspective goes beyond simply considering the transfer of knowledge and skills, emphasizing the importance of how newcomers are integrated into existing team structures.

From a practical standpoint, our findings suggest that organizations should adopt a more strategic approach to talent integration. Rather than focusing solely on an individual's expertise or collaborative skills, managers should consider how new talents can be optimally positioned within both collaboration and knowledge networks. This might involve targeted onboarding processes, mentoring programs, or strategic project assignments that help newcomers build balanced network positions. Furthermore, our research highlights the potential for using network analysis as a tool for innovation management. By mapping and analyzing collaboration and knowledge networks, organizations can identify optimal network structures and intervene to foster more effective knowledge integration and innovation processes.

In conclusion, our study provides a more comprehensive understanding of how talent mobility and network positioning influence team-level exploratory innovation. By highlighting the complex interplay between collaboration and knowledge networks, we contribute to both theoretical discussions on innovation dynamics and practical strategies for talent management in innovative organizations. Future research can build on these findings to further explore the multifaceted relationship between talent mobility, network structures, and organizational innovation.

Limitations and Future Research

While this study provides valuable insights, it is important to acknowledge its limitations and suggest directions for future research. One key limitation is the

narrow scope of talent mobility scenarios examined. Future studies could expand on our findings by investigating a broader range of talent mobility contexts, such as the movement of scientists between research institutions or the transfer of management personnel across organizations. These diverse scenarios could potentially reveal richer and more nuanced innovation mechanisms that occur during talent mobility processes, contributing to a more comprehensive understanding of how different types of talent movement affect innovation dynamics in various organizational settings.

Another important area for future research lies in examining the factors that influence newcomers' evolving positions within collaboration and knowledge networks after joining a team. Future studies could focus on investigating whether and how personal characteristics, collaborative behaviors, or organizational factors affect the trajectory of a newcomer's network centrality. For instance, researchers could explore whether certain personality traits or professional backgrounds are associated with faster integration into central network positions, or how team characteristics and organizational practices influence the rate and extent of newcomers' network position changes. Such research would not only contribute to theoretical knowledge about network dynamics in organizational settings but also provide practical insights for managers seeking to optimize the integration of new talents into their teams.

Acknowledgments

This work was supported by the National Social Science Foundation of China (No. 23&ZD225) and the Data Intelligence and Cross-Innovation Laboratory of Nanjing University.

References

Acar, O. A., Tarakci, M., & Van Knippenberg, D. (2019). Creativity and innovation under constraints: A cross-disciplinary integrative review. *Journal of management*, *45*(1), 96-121.

Acharya, C., Ojha, D., Gokhale, R., & Patel, P. C. (2022). Managing information for innovation using knowledge integration capability: The role of boundary spanning objects. *International Journal of Information Management*, 62, 102438.

Agrawal, A., McHale, J., & Oettl, A. (2017). How stars matter: Recruiting and peer effects in evolutionary biology. *Research Policy*, *46*(4), 853-867.

Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly journal of economics*, *107*(4), 1137-1160.

Bessen, J. (2008). The value of US patents by owner and patent characteristics. *Research Policy*, 37(5), 932-945.

Bower, J. L., & Christensen, C. M. (1996). Disruptive technologies: Catching the wave. *The Journal of Product Innovation Management*, *1*(13), 75-76.

Bunderson, J. S., Van der Vegt, G. S., & Sparrowe, R. T. (2014). Status inertia and member replacement in role-differentiated teams. *Organization Science*, *25*(1), 57-72.

Camarinha-Matos, L. M., & Afsarmanesh, H. (2005). Collaborative networks: a new scientific discipline. *Journal of intelligent manufacturing*, *16*, 439-452.

Cascio, W. F., & Montealegre, R. (2016). How technology is changing work and organizations. *Annual review of organizational psychology and organizational behavior*, *3*(1), 349-375.

Choudhury, P. (2017). Innovation outcomes in a distributed organization: Intrafirm mobility and access to resources. *Organization Science*, *28*(2), 339-354.

Deichmann, D., Moser, C., Birkholz, J. M., Nerghes, A., Groenewegen, P., & Wang, S. (2020). Ideas with impact: How connectivity shapes idea diffusion. *Research policy*, *49*(1), 103881.

Dong, J. Q., & Yang, C. H. (2016). Being central is a double-edged sword: Knowledge network centrality and new product development in US pharmaceutical industry. *Technological Forecasting and Social Change*, *113*, 379-385.

Eslami, H., Ebadi, A., & Schiffauerova, A. (2013). Effect of collaboration network structure on knowledge creation and technological performance: the case of biotechnology in Canada. *Scientometrics*, *97*, 99-119.

Fahrenkopf, E., Guo, J., & Argote, L. (2020). Personnel mobility and organizational performance: The effects of specialist vs. generalist experience and organizational work structure. *Organization Science*, 31(6), 1601-1620.

Ge, C., Huang, K. W., & Kankanhalli, A. (2020). Platform skills and the value of new hires in the software industry. *Research Policy*, *49*(1), 103864.

Guan, J., & Liu, N. (2016). Exploitative and exploratory innovations in knowledge network and collaboration network: A patent analysis in the technological field of nano-energy. *Research policy*, *45*(1), 97-112.

Guan, J., Yan, Y., & Zhang, J. J. (2017). The impact of collaboration and knowledge networks on citations. *Journal of Informetrics*, *11*(2), 407-422.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science*, *308*(5722), 697-702.

Issac, A. C., Bednall, T. C., Baral, R., Magliocca, P., & Dhir, A. (2023). The effects of expert power and referent power on knowledge sharing and knowledge hiding. *Journal of Knowledge Management*, 27(2), 383–403.

Jain, A. (2016). Learning by hiring and change to organizational knowledge: Countering obsolescence as organizations age. *Strategic Management Journal*, 37(8), 1667-1687.

Jain, A., & Huang, K. G. (2022). Learning from the Past: How Prior Experience Impacts the Value of Innovation After Scientist Relocation. *Journal of Management*, 48(3), 571–604.

Jiang, S., Shi, D., & Cheng, Y. (2024). Buyer-initiated knowledge payment facilitates free knowledge contribution in knowledge-sharing platforms: an integration of relational signaling theory and attribution theory. *Journal of Knowledge Management*, *28*(9), 2773-2792.

Jotabá, M. N., Fernandes, C. I., Gunkel, M., & Kraus, S. (2022). Innovation and human resource management: a systematic literature review. *European Journal of Innovation Management*, *25*(6), 1-18.

Kerr, S. P., Kerr, W., Özden, Ç., & Parsons, C. (2016). Global talent flows. *Journal of Economic Perspectives*, *30*(4), 83-106.

Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization science*, *3*(3), 383-397.

Li, V., Mitchell, R., & Boyle, B. (2016). The divergent effects of transformational leadership on individual and team innovation. *Group & Organization Management*, *41*(1), 66-97.

Li, Y., Li, N., Li, C., & Li, J. (2020). The Boon and Bane of Creative "Stars": A Social Network Exploration of How and When Team Creativity Is (and Is Not) Driven by a Star Teammate. *Academy of Management Journal*, 63(2), 613–635.

Lin, R., Lu, Y., Zhou, C., & Li, B. (2022). Rethinking individual technological innovation: Cooperation network stability and the contingent effect of knowledge network attributes. *Journal of Business Research*, *144*, 366-376.

Lingo, E. L. (2023). Digital curation and creative brokering: Managing information overload in open organizing. *Organization Studies*, *44*(1), 105-133.

McAdam, R., & McClelland, J. (2002). Individual and team-based idea generation within innovation management: organisational and research agendas. *European Journal of Innovation Management*, *5*(2), 86-97.

Myers, C. G. (2021). Performance Benefits of Reciprocal Vicarious Learning in Teams. *Academy of Management Journal*, 64(3), 926–947.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, *98*(2), 404-409.

Phelps, C., Heidl, R., & Wadhwa, A. (2012). Knowledge, networks, and knowledge networks: A review and research agenda. *Journal of management*, *38*(4), 1115-1166.

Phelps, C., Heidl, R., & Wadhwa, A. (2012). Knowledge, networks, and knowledge networks: A review and research agenda. *Journal of management*, *38*(4), 1115-1166.

Ren, H., & Zhao, Y. (2021). Technology opportunity discovery based on constructing, evaluating, and searching knowledge networks. *Technovation*, *101*, 102196.

Shi, J., Wang, J., Kang, L., & Sun, J. (2023). How to poach the talents? Role of social capital and contextual knowledge base. *Technological Forecasting and Social Change*, 197, 122905.

Singh, J., & Agrawal, A. (2011). Recruiting for ideas: How firms exploit the prior inventions of new hires. *Management science*, *57*(1), 129-150.

Song, J., Almeida, P., & Wu, G. (2003). Learning–by–hiring: When is mobility more likely to facilitate interfirm knowledge transfer?. *Management science*, *49*(4), 351-365.

Srikanth, K., & Puranam, P. (2014). The firm as a coordination system: Evidence from software services offshoring. *Organization science*, *25*(4), 1253-1271.

Subramaniam, M., & Youndt, M. A. (2005). The influence of intellectual capital on the types of innovative capabilities. *Academy of Management journal*, *48*(3), 450-463.

Tandon, V., Ertug, G., & Carnabuci, G. (2020). How Do Prior Ties Affect Learning by Hiring? *Journal of Management*, 46(2), 287–320.

Tang, T. (Ya), Fang, E. (Er), & Qualls, W. J. (2020). More Is Not Necessarily Better: An Absorptive Capacity Perspective on Network Effects in Open Source Software Development Communities. *MIS Quarterly*, 44(4), 1651–1678.

Tzabbar, D., Cirillo, B., & Breschi, S. (2022). The differential impact of intrafirm collaboration and technological network centrality on employees' likelihood of leaving the firm. *Organization Science*, *33*(6), 2250-2273.

Van der Voet, J., & Steijn, B. (2021). Team innovation through collaboration: How visionary leadership spurs innovation via team cohesion. *Public Management Review*, *23*(9), 1275-1294.

Wang, C., Rodan, S., Fruin, M., & Xu, X. (2014). Knowledge networks, collaboration networks, and exploratory innovation. *Academy of management journal*, *57*(2), 484-514.

Wang, J., & Yang, N. (2019). Dynamics of collaboration network community and exploratory innovation: The moderation of knowledge networks. *Scientometrics*, *121*(2), 1067-1084.

Wang, J., Shi, J., Chen, Y., Kang, L., & Sun, J. (2024). Exploring the impact of inter-organizational knowledge potential difference: an empirical investigation of inventor mobility[J]. *Scientometrics*,2024,129(12):5979–6006.

Wang, M. C., Chen, P. C., & Fang, S. C. (2018). A critical view of knowledge networks and innovation performance: The mediation role of firms' knowledge integration capability. *Journal of Business Research*, *88*, 222-233.

Wang, Q. R., & Zheng, Y. (2022). Nest without birds: Inventor mobility and the left-behind patents. *Research Policy*, 51(4), 104485.

Wang, T., & Zatzick, C. D. (2019). Human Capital Acquisition and Organizational Innovation: A Temporal Perspective. *Academy of Management Journal*, 62(1), 99–116.

Yang, H., Lin, Z., & Peng, M. W. (2011). Behind acquisitions of alliance partners: Exploratory learning and network embeddedness. *Academy of Management Journal*, *54*(5), 1069-1080.

Yang, Y., She, Y., Hong, J., & Gan, Q. (2021). The centrality and innovation performance of the quantum high-level innovation team: the moderating effect of structural holes. *Technology Analysis & Strategic Management*, *33*(11), 1332-1346.

Yuan, Y., & Van Knippenberg, D. (2022). Leader network centrality and team performance: Team size as moderator and collaboration as mediator. *Journal of Business and Psychology*, *37*(2), 283-296.

Zahra, S. A., & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of management review*, *27*(2), 185-203.

Zhang, G. (2021). Employee co-invention network dynamics and firm exploratory innovation: The moderation of employee co-invention network centralization and knowledge-employee network equilibrium. *Scientometrics*, 126(9), 7811–7836.

Author index