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.

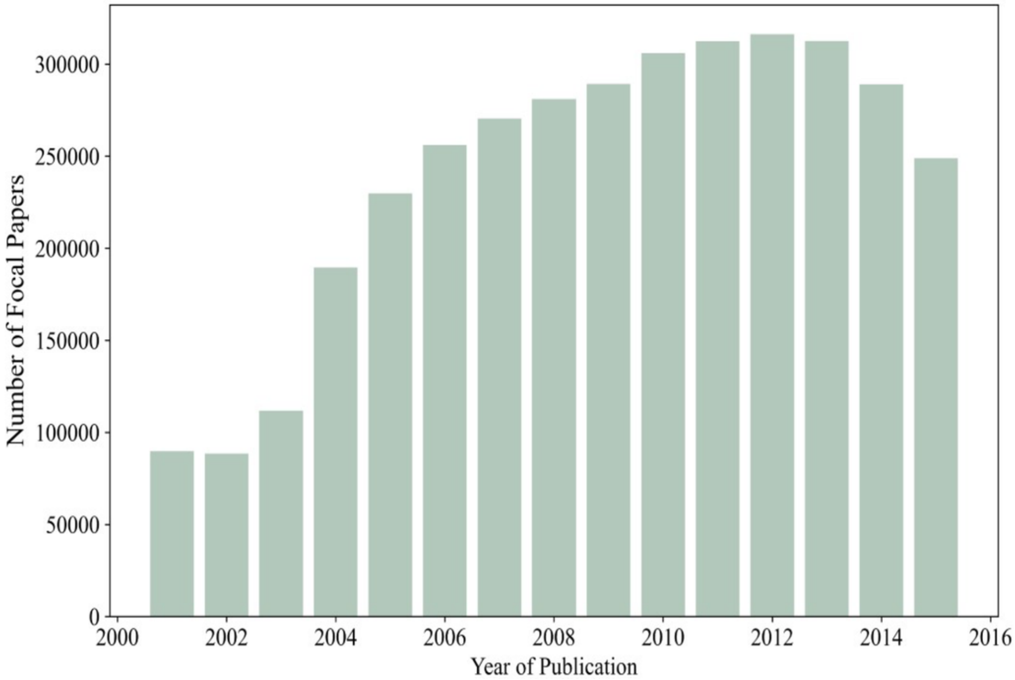

Figure 1. The distribution of the FPs in large-scale dataset from PubMed over years.

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.

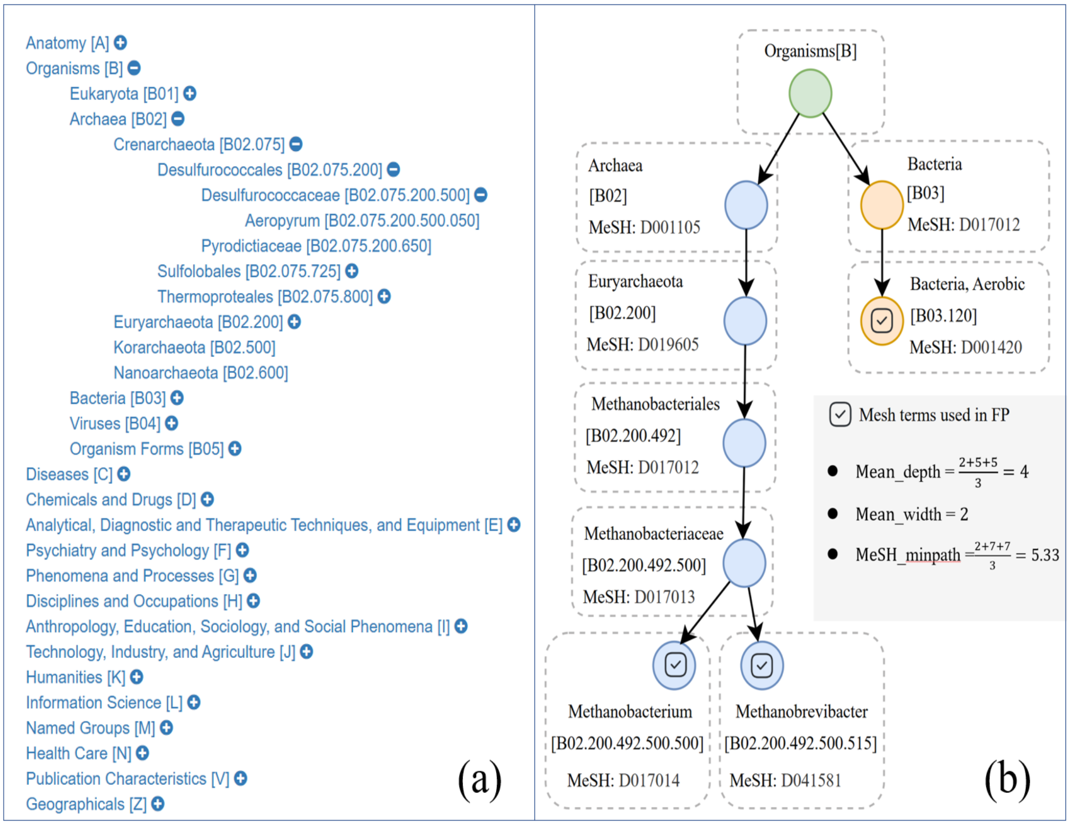
Anatomy [A]
Organisms [B]
Eukaryota [B01]
Archaea [B02]
Crenarchaeota [B02.075]
Desulfurococcales [B02.075.200]
Desulfurococcaceae [B02.075.200.500]
Aeropyrum [B02.075.200.500.050]
Pyrodictiaceae [B02.075.200.650]
Sulfolobales [B02.075.725]
Thermoproteales [B02.075.800]
Euryarchaeota [B02.200]
Korarchaeota [B02.500]
Nanoarchaeota [B02.600]
Bacteria [B03]
Viruses [B04]
Organism Forms [B05]
Diseases [C]
Chemicals and Drugs [D]
Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E]
Psychiatry and Psychology [F]
Phenomena and Processes [G]
Disciplines and Occupations [H]
Anthropology, Education, Sociology, and Social Phenomena [I]
Technology, Industry, and Agriculture [J]
Humanities [K]
Information Science [L]
Named Groups [M]
Health Care [N]
Publication Characteristics [V]
Geographicals [Z]
(a)
Organisms[B]
Archaea [B02] MeSH: D001105
Bacteria [B03] MeSH: D017012
Euryarchaeota [B02.200] MeSH: D019605
Bacteria, Aerobic [B03.120] MeSH: D001420
Methanobacteriales [B02.200.492] MeSH: D017012
Methanobacteriaceae [B02.200.492.500] MeSH: D017013
Methanobacterium [B02.200.492.500.500] MeSH: D017014
Methanobrevibacter [B02.200.492.500.515] MeSH: D041581
☑ Mesh terms used in FP
● Mean_depth = $\frac{2+5+5}{3}$ = 4
● Mean_width = 2
● MeSH_minpath = $\frac{2+7+7}{3}$ = 5.33
(b)

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.

$$Mean_{depth} = \frac{1}{n}\sum_{m=1}^{n} M_d \tag{1}$$

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.

$$Mean\_width\ = Count\ domain(\sum_{m=1}^{n} M_c\ ) \tag{2}$$

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.

$$MeSHmin_{path} = \frac{2 * \sum_{m}^{n} \sum_{m+1}^{n} shortest\ path(p_m, p_{m+1})}{n(n-1)} \tag{3}$$

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.

$$Mean_{year} = \frac{1}{n}\sum_{m=1}^{n}\big(t(m) - t_0(m)\big) \tag{4}$$

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.

$$Sd_{year} = \frac{1}{n}\sum_{m=1}^{n}(a_m - \bar{a})^2 \tag{5}$$

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.

$$U_{test} = \min\left(SumD_{k=1} - \frac{n(n-1)}{2}, SumD_{k=0} - \frac{n(n-1)}{2}\right) \tag{8}$$

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.

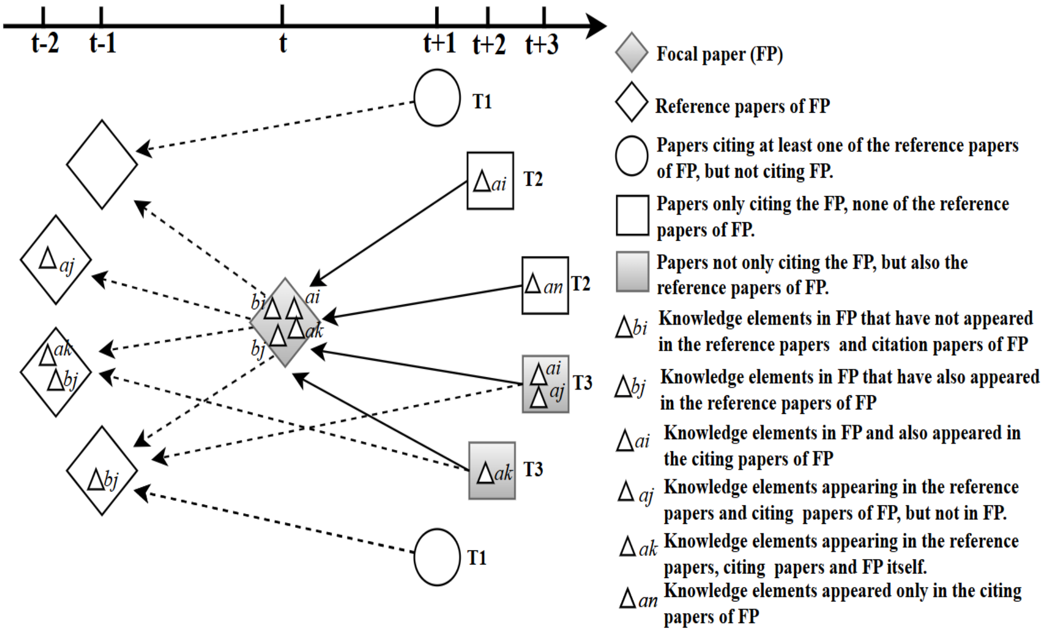
t-2
t-1
t
t+1
t+2
t+3
T1
T2
T2
T3
T3
T1
Δai
Δan
Δai
Δaj
Δak
Δaj
Δak
Δbj
Δbj
bi
ai
ak
bj
Focal paper (FP)
Reference papers of FP
Papers citing at least one of the reference papers of FP, but not citing FP.
Papers only citing the FP, none of the reference papers of FP.
Papers not only citing the FP, but also the reference papers of FP.
Δbi Knowledge elements in FP that have not appeared in the reference papers and citation papers of FP
Δbj Knowledge elements in FP that have also appeared in the reference papers of FP
Δai Knowledge elements in FP and also appeared in the citing papers of FP
Δaj Knowledge elements appearing in the reference papers and citing papers of FP, but not in FP.
Δak Knowledge elements appearing in the reference papers, citing papers and FP itself.
Δan Knowledge elements appeared only in the citing papers of FP

Figure 3. Illustration of the citation pattern with knowledge elements related to FP (Wu et al., 2019; Wang et al., 2023).

$$ED_b = \frac{n_{bi} - n_{bj}}{n_{bi} + n_{bj}} \qquad (9)$$

$$ED_{a,t} = \frac{1}{N}\sum\nolimits_{c=1}^{N} \frac{n_{ai} + n_{an} - n_{aj} - n_{ak}}{n_{ai} + n_{an} + n_{aj} + n_{ak}} \qquad (10)$$

$$ED_t = \beta ED_b + (1-\beta) ED_{a,t} \qquad (11)$$

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.

$$Disruption_i = \alpha_0 + \alpha_1 Mean\_year_i + \alpha_2 Sd\_year_i + \alpha_3 Mean\_depth_i + \alpha_4 Mean\_width_i + \alpha_5 MeSH\_reuse_i + \alpha_6 MeSHmin\_path_i + \alpha_7 Control_i + PY_i + \varepsilon_1 \qquad (12)$$

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.

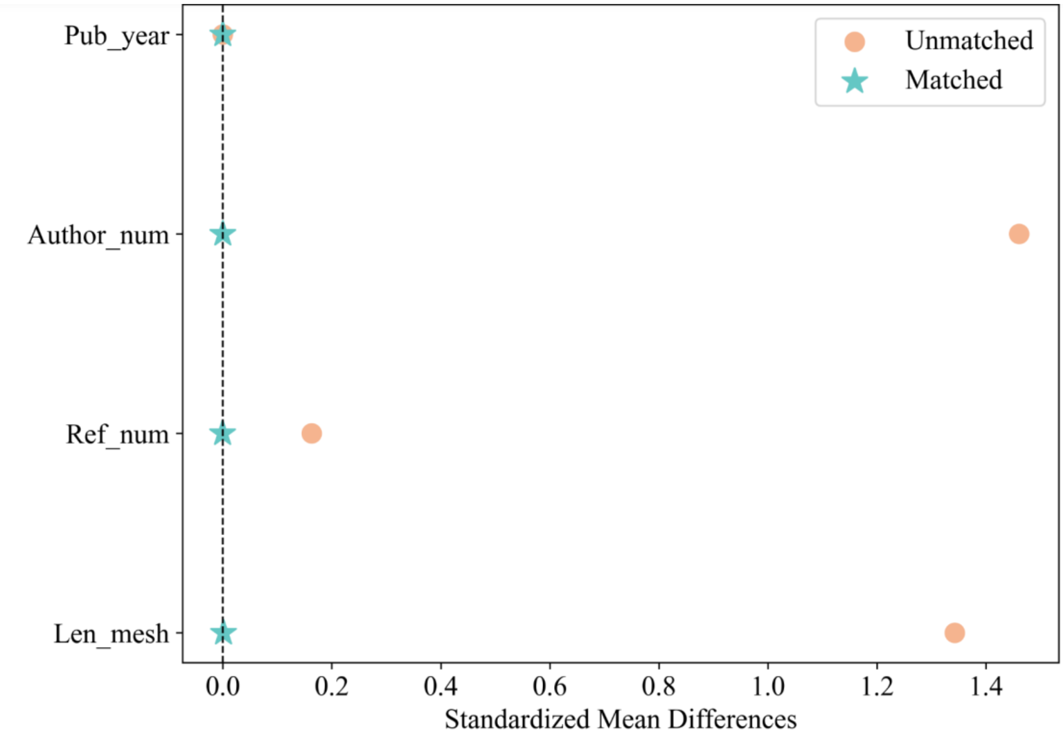

Figure 4. Standardized mean differences of control variables for papers between groups of highly and less disruptive papers before and after EBM.

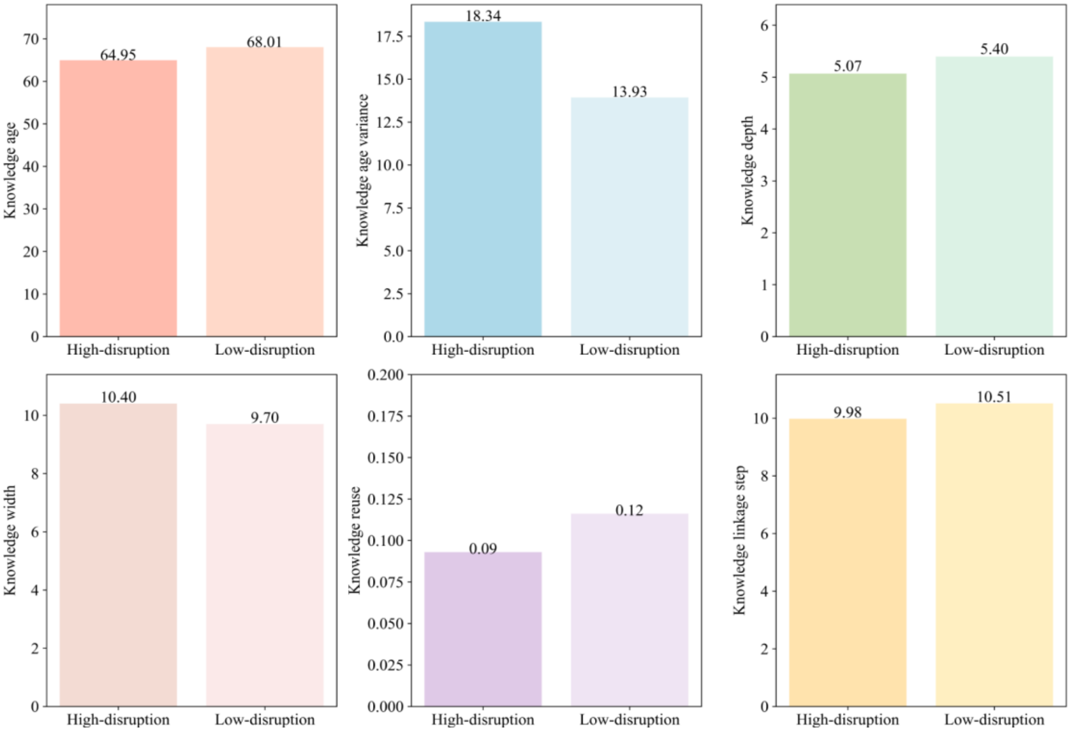

Figure 5. Knowledge features differences of papers between groups of highly and less disruptive papers after EBM.

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.

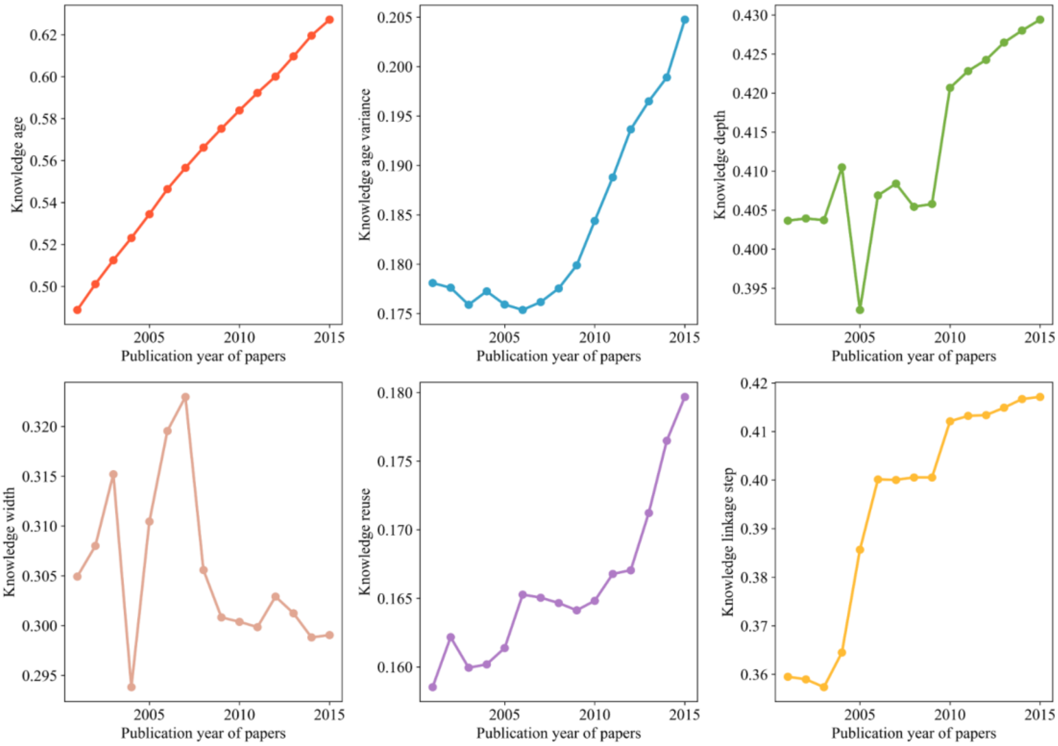

Figure 6. Trends of knowledge features for the publications of Large-scale dataset over years.

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).

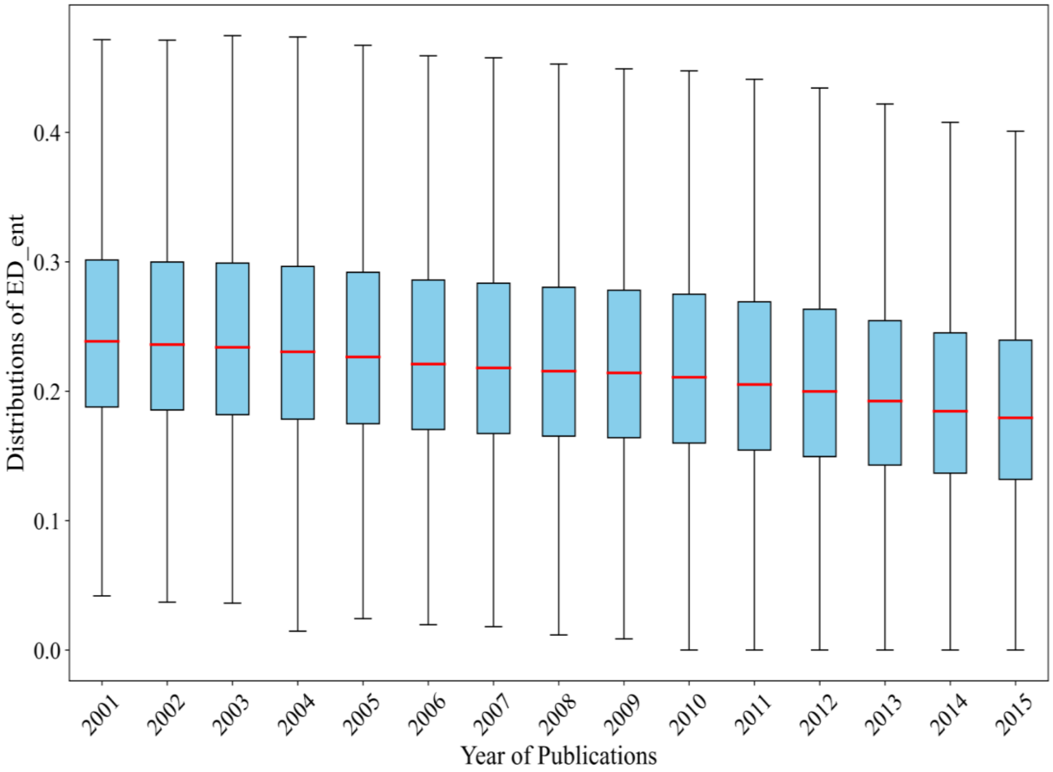

Figure 7. The distribution of disruption scores in the publications of large-scale dataset and their trends over years.

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*** (0.0006)	0.0324*** (0.0005)
Mean_year			-0.0260*** (0.0007)	-0.0835*** (0.0009)
MeSH_reuse			-0.2082*** (0.0004)	-0.1576*** (0.0005)
Pmidmin_path		-0.1926*** (0.0008)		-0.1170*** (0.0009)
Deep_mean		-0.1286*** (0.0008)		-0.1046*** (0.0008)
Wide_mean		0.0213*** (0.0006)		0.0251*** (0.0006)
Len_MeSH	-0.0318*** (0.0004)	-0.0362*** (0.0005)	-0.0721*** (0.0004)	-0.0876*** (0.0005)
Ref_num	-3.3562*** (0.0051)	-3.3208*** (0.0051)	-3.2662*** (0.0051)	-3.3273*** (0.0051)
AuthorNum	-1.8771*** (0.0253)	-1.1085*** (0.0250)	-2.0184*** (0.0246)	-1.6520*** (0.0246)
Pub_year	YES	YES	YES	YES
const	7.1876*** (0.0255)	5.0782*** (0.0260)	6.4454*** (0.0292)	4.1094*** (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.