https://doi.org/10.51408/issi2025_147

Beyond Citations: Tracing and Validating the Rapid Adoption of AlphaFold in Biomedical Research Through Full-Text Analysis

Haochuan Cui[1], Yuzhuo Wang[2], Kai Li[3]

[1] *hcui94@hotmail.com*
School of Computing and Information, University of Pittsburgh, Pittsburgh, PA 15213 (USA)
Knowledge Lab, University of Chicago, 5735 South Ellis Avenue, Chicago, IL 60637 (USA)

[2] *wangyuzhuo@ahu.edu.cn*
School of Management, Anhui University, Hefei 230093 (P.R. China)

[3] *kli16@utk.edu*
School of Information Sciences, University of Tennessee, Knoxville, TN 37996 (USA)

Abstract

The emergence of AlphaFold, a deep learning model for protein structure prediction, has transformed biomedical research. This study analyzes full-text articles from the PubMed Central Open Access dataset to evaluate the dissemination and impact of AlphaFold. Focusing on 8,910 AlphaFold-related articles published between 2018 and 2023, we identify a significant rise in its application across major biomedical fields. Our analysis reveals discrepancies between mentions, citations, and actual usage: citation-based methods capture only 71% of articles mentioning AlphaFold in full text, while half of the articles citing foundational AlphaFold papers do not explicitly reference its name in the citation sentence. Despite being limited by the dataset's scope, this study highlights the need for advanced research methods and infrastructure to accurately assess the impact and usage of AI tools. Future work should explore a broader range of tools and datasets to provide a more comprehensive understanding of AI's influence on scientific research.

Introduction

The modern scientific landscape increasingly relies on advanced information technologies, including artificial intelligence (AI) and deep learning (Gao & Wang, 2024; Stevens et al., 2020). A prominent example of these technologies' transformative impact on science is AlphaFold. Developed by DeepMind in 2020, AlphaFold is a deep learning model designed to predict three-dimensional protein structures based on amino acid sequences (Ruff & Pappu, 2021). Initial testing demonstrated its exceptional accuracy (Jumper et al., 2021; Kovalevskiy et al., 2024), and it has since gained significant traction in fields such as data services, bioinformatics, structural biology, and drug discovery (Varadi & Velankar, 2023). The importance of AlphaFold was further underscored in 2024 when its developers received the Nobel Prize in Chemistry, marking a milestone in recognizing the profound influence of AI technologies on scientific research (Abriata, 2024).

In this project, we aim to comprehensively evaluate the impact of AlphaFold, as an emerging AI technology, on scientific research. AlphaFold provides an ideal case study due to its significant influence, as outlined above. However, assessing the impact of scientific software or algorithmic tools poses substantial challenges. First, these tools may not always be cited or even mentioned in publications. Second, when cited, they may not consistently be represented by the same reference (Li et al., 2019). Consequently, relying solely on citation data to measure the impact of software and algorithms is widely recognized as inadequate (Howison & Bullard, 2016; Wang & Zhang, 2020). We argue that these methodological challenges have important implications for the growing research interest in AI for Science (Stevens et al., 2020). Addressing these issues requires the attention of researchers in scientometrics, research evaluation, and the science of science communities.

This short paper presents preliminary findings aimed at accurately evaluating the impact of AlphaFold. Rather than relying solely on citation data, we utilized full-text academic publications from the PubMed Central Open Access Subset (PMCOA) dataset. By examining the full text of academic publications and analyzing the contexts in which AlphaFold is mentioned, we aim to validate methodologies for tracing the impact of AI tools and develop a more nuanced understanding of how AlphaFold is utilized in scientific research. Specifically, this study seeks to address the following research questions:

RQ1: How has AlphaFold been disseminated in scientific research since its development?

RQ2: In what contexts is AlphaFold used in scientific research?

RQ3: How accurately do citations to AlphaFold papers reflect its impact and usage?

Our findings provide initial empirical evidence of AlphaFold's impact following its development and validate this impact by analyzing the contexts of name mentions. The results highlight the need to distinguish between citations, mentions, and usage of AlphaFold, as significant discrepancies exist among these measures. Furthermore, the findings challenge the validity of using (1) name mentions in titles and abstracts and (2) citations to key AlphaFold publications as proxies for its impact—a common practice in recent research (Hajkowicz et al., 2023; Liu et al., 2021). These insights call for a more nuanced approach to evaluating the influence of AlphaFold and other AI technologies in scientific research.

Methods

This study investigates the impact of AlphaFold on scientific research by analyzing the full-text content of academic papers, by taking the following major steps.

Data Collection

We downloaded a total of 609,615 full-text academic papers from the PubMed Central Open Access (PMCOA) dataset. Following the method proposed by Hsiao and Torvik (2023), we parsed the papers to extract key contextual information for each sentence, including section titles and citation details. Given the distinctiveness of "AlphaFold" as a term in scientific literature, we employed a dictionary-matching approach to identify sentences mentioning AlphaFold, including its key variations such as "AlphaFold" and "Alpha Fold." This process yielded a final dataset with 56,650 sentences from 8,910 papers published between 2018 and 2023 that reference

AlphaFold. To test the accuracy of this approach, we randomly selected 50 sentences from the dataset and 100% of them were the sentences that mentioned AlphaFold.

Entity Feature Identification

From the extracted sentences, we identified the following features for subsequent analysis:

1. **Section of the Sentence**: We analyzed the section of sentences as a signal for understanding how AlphaFold is utilized in scientific research. Previous studies have demonstrated that section titles provide valuable context for identifying the narrative function of sections, particularly within the IMRaD paper structure (Ma et al., 2022). Using a rule-based approach, we categorized section titles into six classical academic sections: Abstract, Introduction, Methods, Results, Discussion, and Others (Sollaci & Pereira, 2004). Keywords used to identify each section is available from our GitHub repository[1].

2. **Narrative Function of the Sentence**: We further leveraged a human-labelled dataset from Jurgens et al. (2018), which includes nearly 2,000 sentences annotated with one of five citation functions: *Uses*, *CompareOrContrast*, *Background*, *Extension*, *Motivation*, and *Future*. The definition of each category is also discussed in our GitHub repository. In this research, we are focused on the category of Uses, as it indicates that AlphaFold is used in the scientific research as a research tool. Using this dataset as training data, we fine-tuned the SciBERT model to classify the narrative function of sentences mentioning AlphaFold in our sample. We split the original dataset into three subsets: 1,600 samples for training, 200 samples for validation, and the remaining samples for testing. The fine-tuned model achieved an F1 score of 76%. To evaluate its performance on AlphaFold-related sentences, we applied the model and randomly selected 20 sentences for testing. Among these, six sentences were classified as 'Use,' and all were correctly identified. The remaining fourteen sentences were correctly classified as 'Not Use,' demonstrating the model's strong ability to distinguish 'Use' from other narrative functions.

3. **Research Areas of the Paper**: We identified the research topics of each paper in the dataset using Medical Subject Headings (MeSH) terms from the PubMed system. Each paper's topics were mapped to one of 16 top-level MeSH categories, representing broad research areas. For instance, the MeSH term "DiGeorge Syndrome," with tree number C16.131.077.019.500, belongs to category C (Diseases). Papers could be associated with multiple research areas.

4. **Representative References of AlphaFold**: We analyzed the references cited by the 8,910 papers mentioning AlphaFold in the PMCOA dataset. Using the PubMed Knowledge Graph (PKG) database, we retrieved all references cited in our final sample. In this preliminary research, we focused on the top three foundational references related to AlphaFold: Jumper et al. (2021), Mirdita et al. (2022), and Varadi et al. (2022).

[1] https://github.com/Wangyuzhuo95/ISSI2025

Results

Rapid Diffusion of AlphaFold in Academic Research

Figure 1(a) illustrates the rapid growth in the number of AlphaFold-related articles and citation sentences since 2018. Furthermore, we identified the top five research areas associated with AlphaFold using the MeSH system: (1) B (Organisms), (2) C (Diseases), (3) D (Chemicals and Drugs), (4) G (Phenomena and Processes), and (5) L (Information Science). Figure 1(b) shows the changing proportions of articles in each of these five areas over time. Notably, the shares of papers in categories D, B, and C have increased, indicating AlphaFold's growing use in applied research. In contrast, the share of papers in category L has declined, suggesting a relatively decreasing focus on analyzing and evaluating AlphaFold in technical literature, such as using AlphaFold to develop other tools and validating AlphaFold.

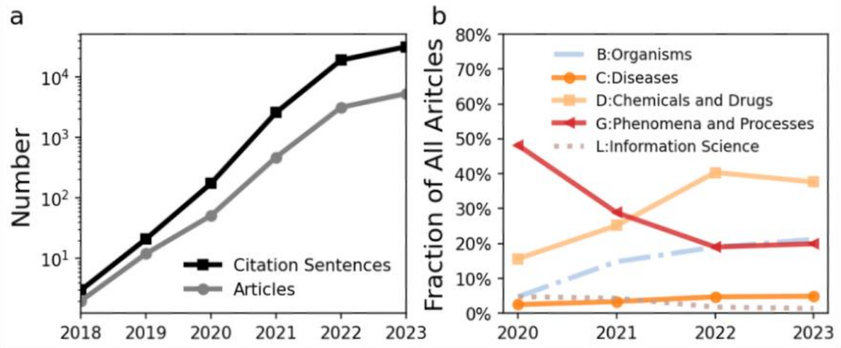


Figure 1. Rapid Diffusion of AlphaFold.

Inaccurate Impact by Traditional Methods

Using full-text academic publications, we compared citations and name mentions of AlphaFold to evaluate the accuracy of tracing its impact. We consider name mentions of AlphaFold as the gold standard for assessing its influence, given the known inaccuracies and limitations of software citations (Li et al., 2019).

Our analysis reveals that, of the 13,396 papers in the whole PMCOA dataset citing at least one of the three foundational references (many of these papers are not in our sample given that they did not mention AlphaFold in the text), only 51.0% explicitly mentioned AlphaFold in the text. Conversely, of the 8,910 papers mentioning AlphaFold, over 2,700 do not cite any of the three references, resulting in an accuracy of 71%. These findings indicate two key points: first, many papers cite key AlphaFold articles for purposes unrelated to AlphaFold, and second, relying solely on citations to trace AlphaFold's impact overlooks many relevant papers.

Figure 2 illustrates the proportion of articles citing the three foundational references (blue line) and those explicitly mentioning AlphaFold (including its variations) in the title or abstract. The proportion number are normalized by overall publication volume. These two measurements correspond to common approaches used to identify publications on the topic. Notably, no papers in our dataset cited the foundational AlphaFold article (Jumper et al., 2021) before its publication in 2021. Additionally, we observe that the trends in both metrics remain consistent across the top five most prominent PMC domain fields, as shown in the supplementary figures.

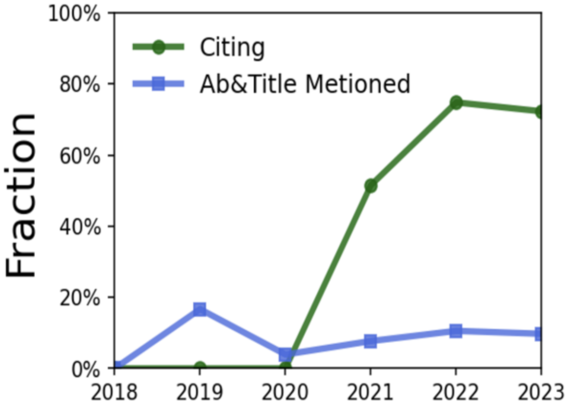
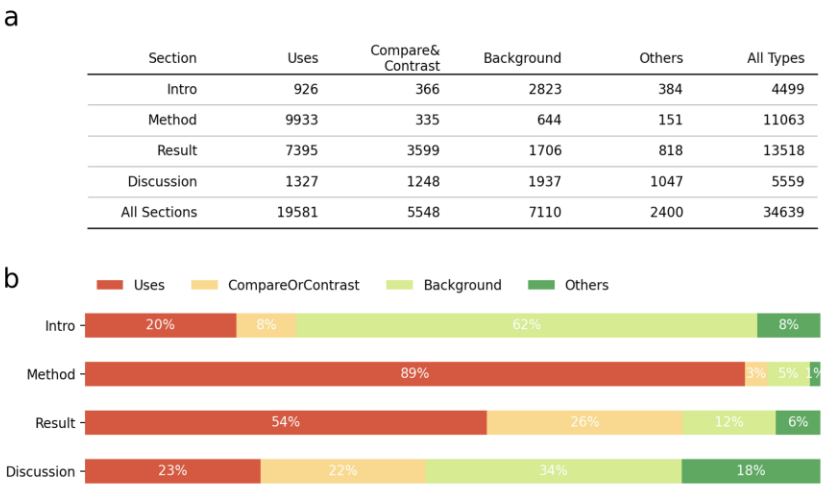


Figure 2. Share of all Articles Citing the Top Three Papers (green line) and Mentioning AlphaFold in the Title or Abstract (blue line).

Our findings carry significant methodological implications for empirical research on the impact of AI on science. Specifically, relying solely on citations or keyword searches in textual fields, such as titles and abstracts, is highly limited. These approaches often fail to capture all relevant articles, overlooking a substantial portion of related research.

For What Purposes is AlphaFold Mentioned in Papers?

To analyze the roles of AlphaFold in academic research, we used a BERT-based model to classify the narrative functions of each sentence into six categories. Figure 3(a) presents the raw counts of sentences for each category, broken down by narrative function and paper section, displaying only the top four categories for clarity. Figure 3(b) illustrates the percentage distribution of narrative functions within each section type. Our findings show that the predominant reason AlphaFold is mentioned in publications is its use in research. Nonetheless, other narrative functions are also frequently represented across the dataset.



a

Section	Uses	Compare& Contrast	Background	Others	All Types
Intro	926	366	2823	384	4499
Method	9933	335	644	151	11063
Result	7395	3599	1706	818	13518
Discussion	1327	1248	1937	1047	5559
All Sections	19581	5548	7110	2400	34639

Figure 3. Distribution of Sentences Across Paper Sections and Narrative Functions.

Figure 4 illustrates the proportion of all articles containing at least one "Use" sentence related to AlphaFold. Our findings indicate that AlphaFold is increasingly utilized as a tool in the corpus. This observation aligns with prior evidence of an "instrumentalization" process for scientific tools within the citation landscape, which can be attributed to the need for such tools to undergo validation before being widely adopted (Li, 2021).

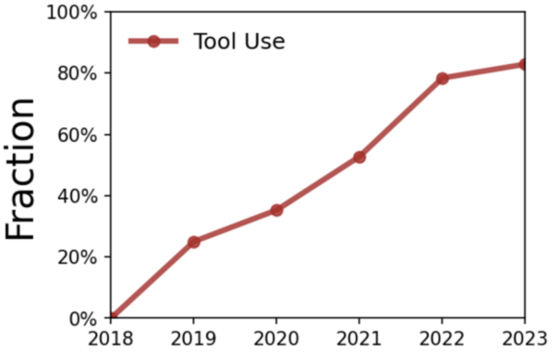


Figure 4. Fraction of Tool Use among AlphaFold-related articles (2018–2023).

Discussions and Conclusion

This paper presents preliminary findings from our project aimed at tracing the impact of AI technologies on science. Our analysis, focused on AlphaFold, highlights the rapid and transformative adoption of this deep learning model in biomedical research, as reflected in the PMCOA corpus. The adoption spans various research fields defined by MeSH terms, with a clear trend toward using AlphaFold in applied research rather than for other technical purposes (such as developing other tools and validating AlphaFold).

A critical insight from our study is the discrepancy between citations, mentions, and actual usage of AlphaFold. Traditional citation analyses often conflate these measures, leading to misunderstandings about the different types of impact associated with software and AI tools. Our findings show that citation-based methods capture 71% of articles mentioning AlphaFold in full text, and only half of the articles citing the three foundational AlphaFold papers explicitly mention AlphaFold within the paper.

These findings carry important implications for scientometrics, research evaluation and science of science research. As AI becomes an indispensable tool for a growing number of researchers, accurately evaluating its impact is an urgent priority for these communities. Our results underscore the significant limitations of relying on citation data and textual queries for assessing the impact of AI tools. These limitations highlight the necessity of full-text analysis for more accurate assessments. While recent studies have leveraged deep learning applications to identify AI technologies in publication texts (Gao & Wang, 2024), building robust data and methodological infrastructures to connect scientific publications to AI tools is essential for advancing this line of research.

In our next steps, we aim to systematically examine usage patterns of other biomedical technologies, such as CRISPR/Cas9. Comparing these patterns with those identified for AlphaFold will provide insights into whether similar trends are shared by other AI tools. This comparative approach will help us develop a more comprehensive understanding of the broader research landscape.

Acknowledgments

This work was partially supported by the National Social Science Fund of China (No.24CTQ027).

References

Abriata, L. A. (2024). The Nobel Prize in Chemistry: past, present, and future of AI in biology. *Communications Biology*, *7*(1), 1409.

Gao, J., & Wang, D. (2024). Quantifying the use and potential benefits of artificial intelligence in scientific research. *Nature human behaviour*, 1-12.

Hajkowicz, S., Sanderson, C., Karimi, S., Bratanova, A., & Naughtin, C. (2023). Artificial intelligence adoption in the physical sciences, natural sciences, life sciences, social sciences and the arts and humanities: A bibliometric analysis of research publications from 1960-2021. *Technology in Society*, *74*, 102260.

Hsiao, T. K., & Torvik, V. I. (2023). OpCitance: Citation contexts identified from the PubMed Central open access articles. *Scientific Data*, 10(1), 243.

Howison, J., & Bullard, J. (2016). Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, *67*(9), 2137-2155.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2018). Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, *6*, 391-406.

Kovalevskiy, O., Mateos-Garcia, J., & Tunyasuvunakool, K. (2024). AlphaFold two years on: Validation and impact. *Proceedings of the National Academy of Sciences*, *121*(34), e2315002121.

Li, K. (2021). The reinstrumentalization of the Diagnostic and Statistical Manual of Mental Disorders (DSM) in psychological publications: A citation context analysis. *Quantitative Science Studies*, *2*(2), 678-697.

Li, K., Chen, P. Y., & Yan, E. (2019). Challenges of measuring software impact through citations: An examination of the lme4 R package. *Journal of Informetrics*, *13*(1), 449-461.

Liu, N., Shapira, P., & Yue, X. (2021). Tracking developments in artificial intelligence research: constructing and applying a new search strategy. *Scientometrics*, *126*(4), 3153-3192.

Ma, B., Zhang, C., Wang, Y., & Deng, S. (2022). Enhancing identification of structure function of academic articles using contextual information. *Scientometrics*, 127(2), 885-925.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: making protein folding accessible to all. *Nature methods*, *19*(6), 679-682.

Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and implications for intrinsically disordered proteins. *Journal of molecular biology*, *433*(20), 167208.

Sollaci LB, Pereira MG. (2004). The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *J Med Libr Assoc*, 92(3), 364-7.

Stevens, R., Taylor, V., Nichols, J., Maccabe, A. B., Yelick, K., & Brown, D. (2020). *AI for science: Report on the department of energy (doe) town halls on artificial intelligence (ai) for science* (No. ANL-20/17). Argonne National Lab. (ANL), Argonne, IL (United States).

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, *50*(D1), D439-D444.

Varadi, M., & Velankar, S. (2023). The impact of AlphaFold Protein Structure Database on the fields of life sciences. *Proteomics*, *23*(17), 2200128.

Wang, Y., & Zhang, C. (2020). Using the full-text content of academic articles to identify and evaluate algorithm entities in the domain of natural language processing. *Journal of informetrics*, *14*(4), 101091