Guo *et al. Molecular Cancer* (2024) 23:223
https://doi.org/10.1186/s12943-024-02140-6

Molecular Cancer

CORRESPONDENCE Open Access

Artificial intelligence alphafold model for molecular biology and drug discovery: a machine-learning-driven informatics investigation

Song-Bin Guo[1,2*†], Yuan Meng[2†], Liteng Lin[3†], Zhen-Zhong Zhou[1,2], Hai-Long Li[1,2], Xiao-Peng Tian[1,2*] and Wei-Juan Huang[4*]

Abstract

AlphaFold model has reshaped biological research. However, vast unstructured data in the entire AlphaFold field requires further analysis to fully understand the current research landscape and guide future exploration. Thus, this scientometric analysis aimed to identify critical research clusters, track emerging trends, and highlight underexplored areas in this field by utilizing machine-learning-driven informatics methods. Quantitative statistical analysis reveals that the AlphaFold field is enjoying an astonishing development trend (Annual Growth Rate = 180.13%) and global collaboration (International Co-authorship = 33.33%). Unsupervised clustering algorithm, time series tracking, and global impact assessment point out that Cluster 3 (Artificial Intelligence-Powered Advancements in AlphaFold for Structural Biology) has the greatest influence (Average Citation = 48.36 ± 184.98). Additionally, regression curve and hotspot burst analysis highlight "structure prediction" (s = 12.40, $R^2 = 0.9480$, $p = 0.0051$), "artificial intelligence" (s = 5.00, $R^2 = 0.8096$, $p = 0.0375$), "drug discovery" (s = 1.90, $R^2 = 0.7987$, $p = 0.0409$), and "molecular dynamics" (s = 2.40, $R^2 = 0.8000$, $p = 0.0405$) as core hotspots driving the research frontier. More importantly, the Walktrap algorithm further reveals that "structure prediction, artificial intelligence, molecular dynamics" (Relevance Percentage[RP] = 100%, Development Percentage[DP] = 25.0%), "sars-cov-2, covid-19, vaccine design" (RP = 97.8%, DP = 37.5%), and "homology modeling, virtual screening, membrane protein" (RP = 89.9%, DP = 26.1%) are closely intertwined with the AlphaFold model but remain underexplored, which implies a broad exploration space. In conclusion, through the machine-learning-driven informatics methods, this scientometric

†Song-Bin Guo, Yuan Meng, and Liteng Lin are the joint first authors of this work.

*Correspondence:
Song-Bin Guo
guosb.sysucc@yeah.net
Xiao-Peng Tian
tianxp@sysucc.org.cn
Wei-Juan Huang
wjhuang@jnu.edu.cn

Full list of author information is available at the end of the article

© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

analysis offers an objective and comprehensive overview of global AlphaFold research, identifying critical research clusters and hotspots while prospectively pointing out underexplored critical areas.

Keywords AlphaFold, Bibliometrics, Artificial intelligence, Molecular dynamics, Structure prediction, Drug discovery

Introduction

The advent of AlphaFold has revolutionized structural biology, marking a new era in protein structure prediction with unprecedented accuracy [1]. Since its inception, the AlphaFold model has provided invaluable insights into the complex structure of biomolecules, rapidly serving as a cornerstone of molecular biology, genomics, and drug discovery [1–3].

As a cross-disciplinary subject, scientometrics takes scientific knowledge carriers and their internal related information as a data source and utilizes mathematical and statistical methods to qualitatively and quantitatively analyze their quantity, distribution, structure, and evolution, ultimately providing solid evidence for future research strategies and investment decisions in a specific field [4, 5]. Scientometric analysis is widely wielded not only in computer science, but more recently in many medical domains, including basic medicine, and clinical medicine [6–8].

Recently, research related to AlphaFold has experienced explosive growth, resulting in the accumulation of vast amounts of unstructured data. To date, the overall research landscape, evolving trends, and future development of the AlphaFold field remain unclear. The machine-learning-driven scientometric analysis coincides with a significant opportunity for scholars in the AlphaFold field to address these challenges in a short time.

Therefore, through the machine-learning-driven informatics methods, this scientometric analysis aimed to identify critical research clusters, track emerging trends, and highlight underexplored critical areas that hold promise for future exploration.

Materials and methods

Data source

Several biomedical databases exist in scientometrics research, including Web of Science (WOS), Scopus, PubMed, MEDLINE, and Embase. Although integrating different databases can provide more information, there is a large amount of duplicated data between different databases. Combining data from different sources may introduce unnecessary confounding factors affecting the data quality and, ultimately, the experimental conclusions. Choosing a single, high-quality, authoritative database as a data source is often used as a routine strategy for scientometrics analysis. This not only represents the main trends and directions of the entire research field but also minimizes confounding factors and ensures data quality. Due to its comprehensiveness and authority, the WOS core database was chosen as the data source for this informatics analysis study.

Data gathering and sample size

The advanced search function of the WOS database was used for data filtering during the data collection process for the informatics analysis. Firstly, this study utilized the search formula TS=(AlphaFold*) to obtain all AlphaFold related studies ($n=1818$). The time frame was from January 1, 2019, to May 28, 2024. Further, we removed non-peer-reviewed ($n=132$) and non-English ($n=6$) documents. Finally, all eligible records were exported as plain text format files. The raw data ($n=1680$) was exported on May 28, 2024.

Unsupervised clustering algorithm

Unsupervised clustering algorithms do not require prelabeled data, making them particularly useful for exploratory analysis, especially when dealing with data that lacks clearly defined category labels. Research hotspot analysis often involves large amounts of unlabeled data with no predefined categories, so unsupervised clustering algorithms can automatically discover structures and patterns based on the inherent features of the data, making them especially suitable for clustering research hotspots within the AlphaFold field. What's more, Unsupervised clustering algorithms can identify natural groups or patterns within the data, which is particularly important in research hotspot analysis. Through clustering, it is possible to identify research directions or topics that naturally form within the AlphaFold field. We used VOSviewer to conduct unsupervised clustering analysis [9, 10]. Specifically, we extracted 4268 keywords from 1680 studies as objects (representing research topic). In the dataset, the frequency of co-occurrence between two objects (representing research topic) was used to represent the link strength between them. The total link strength of an object is the sum of its link strengths with all other objects in the dataset. These co-occurrence relationships are represented in a co-occurrence matrix, where each element reflects the link strength between pairs of objects. We then applied the Louvain algorithm to the co-occurrence matrix for clustering analysis, grouping data objects with similar co-occurrence patterns into the same cluster.

Calculating average publication year and average citation across six clusters

We used VOSviewer to identify 6 clusters among the 1680 studies based on the keywords. Table S2 shows detailed information on the top twelve research hotspots of the six clusters. By calculating the mean and standard deviation of the publication year and citation of research hotspots in each cluster, the average publication year and standard deviation, average citation and standard deviation of each cluster are obtained.

Hotspot burst analysis

The hotspot burst analysis was conducted to identify and quantify the time period when a research hotspot receives a significant increase in attention, indicating the degree of concentration of a particular research hotspot over a period of time. This analysis is critical to understanding temporal trends in research field and identifying emerging topics that may deserve further exploration. We use the R package "bibliometrix" for hotspot burst analysis. "bibliometrix" is well suited for bibliometric studies and provides a powerful tool for temporal analysis of research topics [11]. The input data consisted of 4268 keywords. The parameters were set as follows: Word Minimum Frequency set to 4 and Number of Words per Year set to 3. Finally, the results were visualized using the R package "ggplot2".

Regression analysis

Regression analysis is a classical statistical method used primarily to quantify the relationship between two or more variables. In this study, we aimed to understand the relationship between the frequency of occurrence of a particular research topic and time. Therefore, regression analysis was chosen to determine whether there is a linear relationship between the two and to assess the strength of this relationship. We used the R package "bibliometrix" to extract 4268 keywords from 1680 studies and counted the frequency of these keywords across years, focusing on the top 100 keywords with the highest overall frequency. Subsequently, we conducted a regression analysis to analyze the correlation between the frequency of keyword occurrence and time. P-values less than 0.05 indicate that the correlation is statistically significant, "R^2" denotes the coefficient of determination, and "s" denotes the slope of the fitted curve.

Walktrap-based community analysis of alphafold research

Common community detection algorithms include the Walktrap algorithm, Girvan-Newman algorithm, Louvain algorithm, Label Propagation algorithm, Spectral Clustering algorithm, Infomap algorithm and so on. Each algorithm has its unique advantages and suitable application scenarios. For example, the Girvan-Newman algorithm is well suited for smaller networks, the Louvain algorithm performs well in large-scale networks, and the Infomap algorithm has an advantage in networks with multilayer structures. In this study, we chose the Walktrap algorithm for the following main reasons: Adaptation to Complex Networks: The Walktrap algorithm effectively handles complex and densely connected networks, such as those in the AlphaFold research field, capturing the deep relationships between research topics; Natural Cluster Identification: Through the process of random walks, the algorithm can identify natural clusters among research topics, making it ideal for analyzing dynamically evolving research hotspots and trends; Ease of Interpretation: The results of the Walktrap algorithm are easy to interpret, making it straightforward for researchers to understand the data structure and research findings.

The Walktrap algorithm is a community detection algorithm based on random walks, used to identify communities or clusters (representing research topics) within networks. The algorithm simulates random walks between nodes (representing research topics) in the network to capture the close relationships between nodes, thereby identifying clusters of closely related nodes. Each node (representing research topic) in the network is initially considered as an independent community. The distance between different communities is calculated based on the similarity between research topics. Briefly, the distance matrix was constructed by co-occurrence. The co-occurrence frequency was used as a proxy for similarity, where higher co-occurrence frequencies indicated greater similarity. These similarity scores were then inverted to form the distance matrix, with lower distance values indicating higher similarity between topics. The two communities with the shortest distance are merged, gradually reducing the number of communities. The algorithm selects the partition that best defines the community structure, meaning tight connections within communities and sparse connections between communities.

Results

We comprehensively collected all studies related to AlphaFold to date. To ensure the quality of the included research, we excluded non-peer-reviewed and non-English articles. Statistical results show that AlphaFold is now a burgeoning field. Since AlphaFold first debuted at CASP13, the number of related peer-reviewed English studies has surged to 1680, with an annual growth rate of 180.13% (Table S1). The proportion of international co-authorship has reached 33.33% (Table S1), highlighting the trend towards global collaborative research in the AlphaFold field. To investigate the major research clusters in the AlphaFold field and conduct subsequent time-series and global impact analyses, we applied an

unsupervised clustering algorithm to cluster global research hotspots and identified six clusters: Cluster 1 (Applications of AlphaFold-Based Protein Prediction in Virology and Immunology), Cluster 2 (Application of AlphaFold in Gene Mutation and Gene Expression Regulation), Cluster 3 (Artificial Intelligence[AI]-Powered Advancements in AlphaFold for Structural Biology), Cluster 4 (AlphaFold's Role in Drug Discovery and Molecular Dynamics), Cluster 5 (AlphaFold-Driven Enzyme Engineering and Mechanistic Insights), and Cluster 6 (AlphaFold in Disease-Related Structural Biology and Genomic Research) (Fig. 1A; Table S2). An in-depth analysis of the time series and global impact was further conducted to track the development of the six clusters in the AlphaFold field. The results reveal that Cluster 3, the earliest emerging key cluster (Average Publication Year=2022.72±0.40), has demonstrated remarkable influence and importance (Average Citation=48.36±184.98). In contrast, cluster 5, as an emerging force, is just beginning to show potential and has significant room for growth (Average Publication Year=2022.93±0.31; Average Citation=7.29±19.25) (Fig. 1B). Spatial density networks driven by total link strength (TLS) or occurrence frequency (OF) further provide comprehensive and intuitive visual insights into AlphaFold research hotspots (Fig. 1C and D).

The subsequent regression curve analysis, aimed at exploring trends in research topics, reveals a strong upward trajectory in areas such as: "structure prediction" (s=12.40 [95% CI: 7.062, 17.74], R^2=0.9480, p=0.0051), "AI" (s=5.00 [95% CI: 0.5446, 9.455], R^2=0.8096, p=0.0375), "drug discovery" (s=1.90 [95% CI: 0.1472, 3.653], R^2=0.7987, p=0.0409), and "molecular dynamics" (s=2.40 [95% CI: 0.1951, 4.605], R^2=0.8000, p=0.0405) (Fig. 2A). Hotspot burst analysis was employed to detect sudden bursts in AlphaFold research, pinpointing topics that have seen rapid growth over a short period. This analysis highlights that "protein structure", "molecular

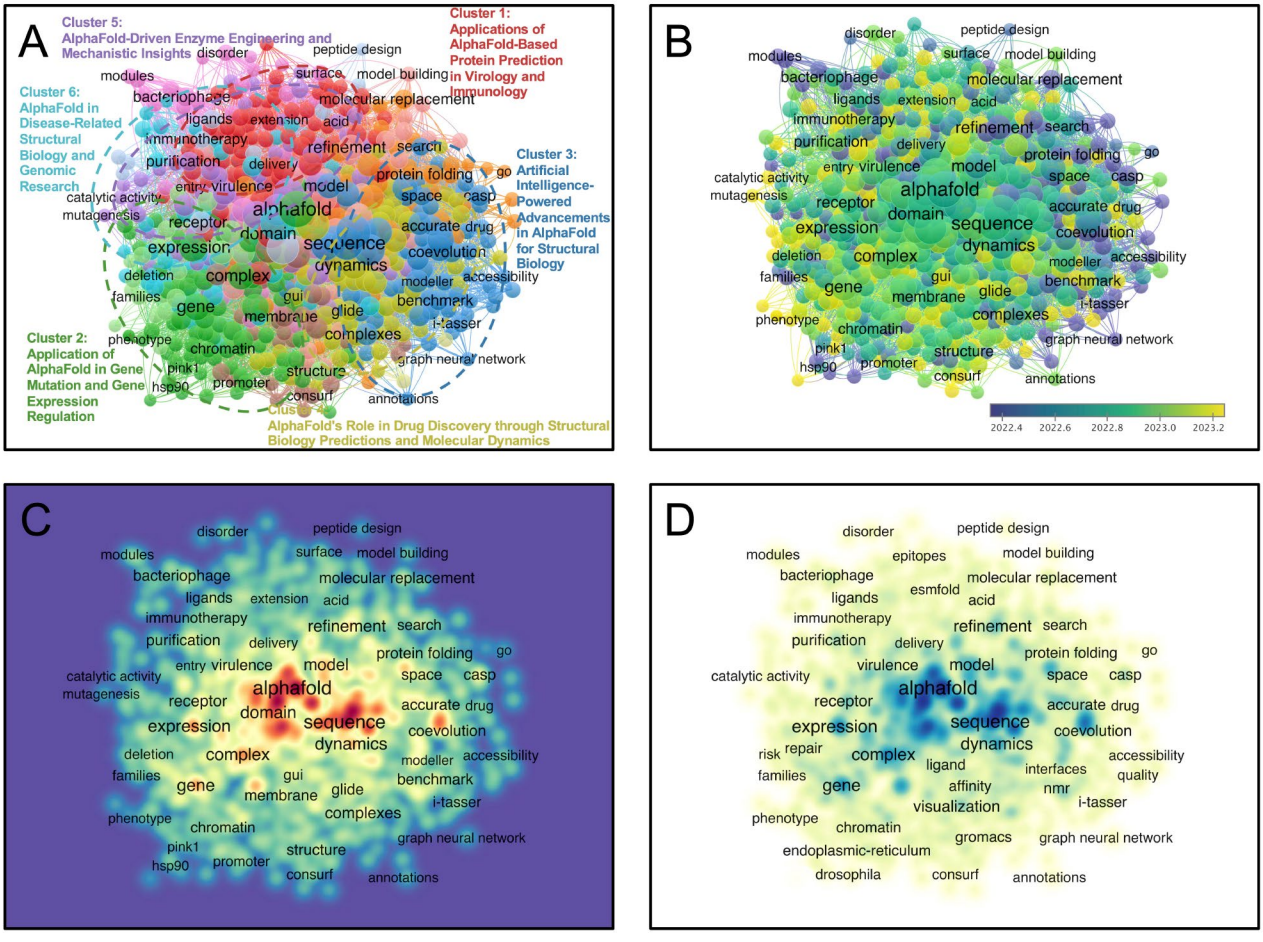

Fig. 1 A concise overview of the spatial and temporal distribution of high-quality research hotspots in the AlphaFold domain. (**A**) Machine learning-based unsupervised clustering algorithm divides global research hotspots into six major clusters; (**B**) Time series tracking reveals the global temporal distribution pattern of research hotspots; (**C**) Spatial density network graph visualizes the total link strength (TLS) of global research hotspots (**D**) Spatial density network graph visualizes the occurrence frequency (OF) of global research hotspots

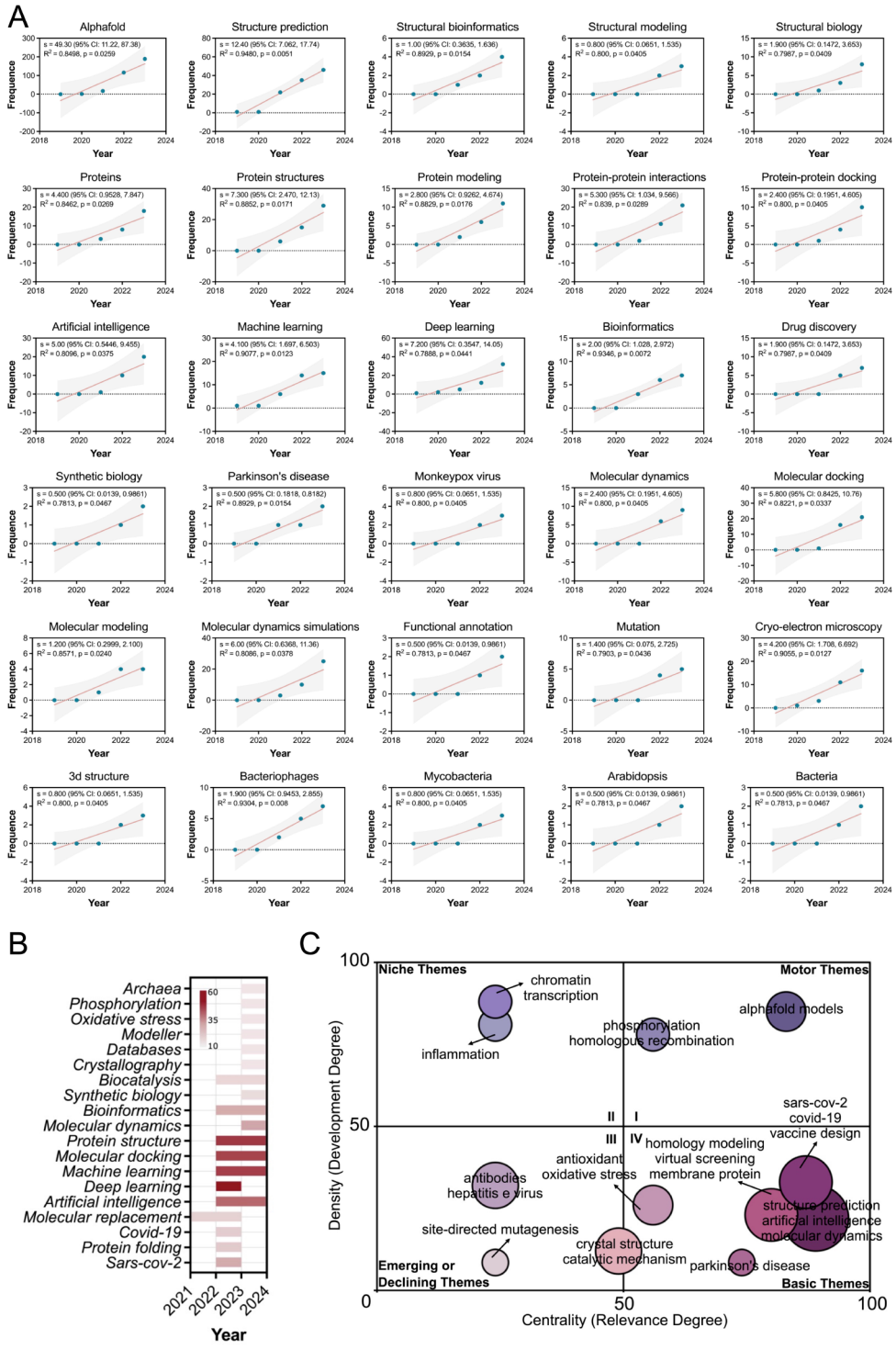

Fig. 2 Development trends and future directions in the AlphaFold domain. (**A**) Regression curve analysis reveals the correlation between the annual occurrence frequency of research hotspots and publication time. 's' represents the slope of the fitted curve, 'R^2' represents the coefficient of determination, and 'CI' represents the confidence interval. (**B**) Hotspot burst analysis for global research hotspots of AlphaFold model. (**C**) The graph-based community detection algorithm Walktrap reveals the critical but still underdeveloped directions for this field that deserve further research. The X-axis represents the relevance to the field, and the Y-axis represents the relative development degree

dynamics", and "AI" are emerging as prominent focuses for future research (Fig. 2B).

Notably, the community detection and network analysis algorithm named Walktrap was used to uncover relationships between research hotspots and identify areas closely related to the AlphaFold that remain underexplored. The results indicate that "structure prediction", "AI", and "molecular dynamics" are closely intertwined with AlphaFold, as indicated by a high Relevance Percentage (RP = 100%). However, these areas also exhibit a relatively low Development Percentage (DP = 25.0%), suggesting that while they are central to the AlphaFold field, they have still not been fully explored. This combination of high relevance and low maturity underscores a broad exploration space and significant research potential, indicating substantial opportunities for future advancements and breakthroughs in these areas (Fig. 2C).

Discussion

This study employs a rigorous scientometric analysis to illuminate the landscape of global AlphaFold research. By identifying critical research clusters and hotspots, it underscores the current trends in the entire AlphaFold field while also revealing underexplored areas that hold significant promise for future investigation.

Multiple informatics analyses indicate that AI is the most influential direction in the AlphaFold field, supported by its highest average citation rate (Average Citation = 48.36 ± 184.98). Specifically, in the global impact analysis of the six research clusters (Fig. 1A; Table S2), Cluster 3 (AI-Powered Advancements in AlphaFold for Structural Biology) has an average citation of 48.36 ± 184.98; while the average citation of Cluster 1 = 5.57 ± 7.74; Cluster 2 = 5.32 ± 9.38; Cluster 4 = 13.13 ± 26.68; Cluster 5 = 7.29 ± 19.25; Cluster 6 = 3.55 ± 4.21. Comparative analysis reveals that Cluster 3 has the highest average citation among all the clusters, thus demonstrating that the AI direction is the most influential. Additionally, AI is a promising direction for future development, as it plays a central role (RP = 100%) but remains underexplored (DP = 25.0%). In comparison, other research directions show statistics such as "sars-cov-2, covid-19, vaccine design" (RP = 97.8%, DP = 37.5%) and "homology modeling, virtual screening, membrane protein" (RP = 89.9%, DP = 26.1%). These findings suggest that other research directions either have lower relevance or offer less room for development than AI direction. Therefore, after a comprehensive comparison, this study conclude that AI is a promising direction for future development in the AlphaFold field.

Each breakthrough of AlphaFold has been deeply marked by the integration and advancement of AI technology, serving as its internal engine of continuous innovation [1–3]. Although the latest AlphaFold 3 has significantly improved prediction accuracy, the extreme complexity of biological systems and the high diversity of protein structures pose substantial challenges in accurately simulating protein dynamics and interaction mechanisms under various environmental conditions [2, 12]. This underscores the urgent need for AI models capable of handling complex scenarios with higher levels of intelligence. Additionally, the development and training of AI models heavily rely on large-scale, diverse datasets [13]. Therefore, expanding the coverage and representativeness of datasets, enhancing computational efficiency, and reducing resource consumption are key issues that need to be addressed in the AI area. As the cornerstone of AlphaFold's development, the continuous optimization and upgrading of AI technology will undoubtedly inject continuous momentum into AlphaFold's future exploration, leading it to new heights.

Interestingly, the Walktrap algorithm and regression analysis indicate that "membrane protein" (RP = 98.9%, DP = 26.1%) and "drug discovery" (s = 1.90 [95% CI: 0.1472, 3.653], $R^2 = 0.7987$, $p = 0.0409$) are closely linked with AlphaFold, yet these areas remain underdeveloped. As we all know, membrane proteins are involved in various critical biological processes, including signal transduction, ion transport, and cell adhesion. They represent a significant portion of drug targets due to their roles in cellular communication and metabolism [14]. Determining the structure of membrane proteins is particularly challenging due to their amphipathic nature and the difficulty in crystallizing them [15]. AlphaFold's ability to predict these complex structures addresses a significant gap in structural biology [3, 12]. Despite AlphaFold's remarkable potential in advancing membrane protein structure resolution and drug design, its accuracy in predicting protein-protein interactions and capturing the dynamic conformational changes requires improvement [16, 17]. For example, G protein-coupled receptors (GPCRs) are classical membrane proteins and are widely used as drug targets. While AlphaFold2 can predict the static structures of GPCRs, it struggles to capture the dynamic conformational changes that occur when GPCRs interact with ligands [12]. This key limitation remains with the newly introduced AlphaFold3, despite algorithmic improvements. Similarly, E3 ubiquitin ligases typically adopt an open conformation in the absence of a ligand and transition to a closed conformation upon ligand binding. However, AlphaFold3 still fails to accurately predict these conformational shifts across different states [2]. To address this limitation, future research could focus on developing more efficient algorithms that incorporate time-series data, reinforcement learning, and molecular dynamics simulations, which could significantly enhance AlphaFold's ability to predict protein dynamics and conformational diversity. (Supplementary file 3).

In conclusion, AlphaFold has undoubtedly injected significant momentum into the areas of membrane protein science and drug discovery, but it also faces numerous challenges that require further exploration. With continuous innovations and iterations in artificial intelligence technology, we believe that AlphaFold will demonstrate even greater performance in these domains in the future.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12943-024-02140-6.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Acknowledgements
We sincerely acknowledge the time and effort the editors and reviewers put into our manuscript.

Author contributions
Conceived and designed the experiments: WJH, XPT, and SBG; Performed the experiments: SBG, YM, LTL, ZZZ, and HLL; Analyzed and interpreted the data: SBG, YM, and LTL; Wrote the paper: SBG, YM, LTL, WJH, and XPT; Administered and supervised the project: WJH and XPT. All authors have read and agreed to the final version of the manuscript.

Funding
This work was supported by grants from National Natural Science Foundation of China (82422010, 82370190 to X.-P. Tian), and Guangdong Basic and Applied Basic Research Foundation (2024B1515020026 to X.-P. Tian; 2024A1515010185, 2023A1515012282 to W.-J. Huang).

Data availability
No datasets were generated during the current study.

Declarations

Ethics approval and consent to participate
The study did not involve humans or animals, so it did not require ethical approval from institutional committees or consent from patient participation.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of Medical Oncology, Sun Yat-Sen University Cancer Center, Guangzhou 510060, P. R. China
[2]State Key Laboratory of Oncology in South China, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, P. R. China
[3]Laboratory of Interventional Radiology, Department of Minimally Invasive Interventional Radiology and Interventional Cancer Center, The Second Affiliated Hospital, Guangzhou Medical University, Guangzhou 510260, P. R. China
[4]Department of Pharmacology, College of Pharmacy, Jinan University, Guangzhou 510632, P. R. China

Received: 3 September 2024 / Accepted: 30 September 2024
Published online: 05 October 2024

References

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
2. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.
3. Yang Z, Zeng X, Zhao Y, Chen R. AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduct Target Ther. 2023;8:115.
4. Donthu N, Kumar S, Mukherjee D, Pandey N, Lim WM. How to conduct a bibliometric analysis: an overview and guidelines. J Bus Res. 2021;133:285–96.
5. Guo S-B, Du S, Cai K-Y, Cai H-J, Huang W-J, Tian X-P. A scientometrics and visualization analysis of oxidative stress modulator Nrf2 in cancer profiles its characteristics and reveals its association with immune response. Heliyon. 2023;9:e17075.
6. Emmert-Streib F, Tripathi S, Dehmer M. Analyzing the Scholarly Literature of Digital Twin Research: Trends, Topics and structure. IEEE Access. 2023;11:69649–66.
7. Guo S-B, Feng X-Z, Huang W-J, Zhou Z-Z, Tian X-P. Global research hotspots, development trends and prospect discoveries of phase separation in cancer: a decade-long informatics investigation. Biomark Res. 2024;12:39.
8. Guo S-B, Hu L-S, Huang W-J, Zhou Z-Z, Luo H-Y, Tian X-P. Comparative investigation of neoadjuvant immunotherapy versus adjuvant immunotherapy in perioperative patients with cancer: a global-scale, cross-sectional, large-sample informatics study. International Journal of Surgery [Internet]. 2024 [cited 2024 Aug 31]; https://journals.lww.com/https://doi.org/10.1097/JS9.0000000000001479
9. Van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. Scientometrics. 2017;111:1053–70.
10. Guo S-B, Pan D-Q, Su N, Huang M-Q, Zhou Z-Z, Huang W-J, et al. Comprehensive scientometrics and visualization study profiles lymphoma metabolism and identifies its significant research signatures. Front Endocrinol. 2023;14:1266721.
11. Aria M, Cuccurullo C. Bibliometrix: an R-tool for comprehensive science mapping analysis. J Informetrics. 2017;11:959–75.
12. Wang L, Wen Z, Liu S-W, Zhang L, Finley C, Lee H-J, et al. Overview of AlphaFold2 and breakthroughs in overcoming its limitations. Comput Biol Med. 2024;176:108620.
13. Wang H, Fu T, Du Y, Gao W, Huang K, Liu Z, et al. Scientific discovery in the age of artificial intelligence. Nature. 2023;620:47–60.
14. Tosaka T, Kamiya K. Function investigations and applications of membrane proteins on Artificial lipid membranes. Int J Mol Sci. 2023;24:7231.
15. Liu S, Li S, Krezel AM, Li W. Stabilization and structure determination of integral membrane proteins by termini restraining. Nat Protoc. 2022;17:540–65.
16. Karelina M, Noh JJ, Dror RO. How accurately can one predict drug binding modes using AlphaFold. Models? Elife. 2023;12:RP89386.
17. Callaway E. AlphaFold found thousands of possible psychedelics. Will its predictions help drug discovery? Nature. 2024;626:14–5.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.