Postprint of the New Paradigm for AI-Driven Life Sciences Research

Authors: Li Xin, Yu Hanchao

Date: 2024-03-27T00:00:00+00:00

Abstract

The rapid development of biotechnology and information technology has ushered life sciences into a new era of data explosion, wherein traditional research paradigms struggle to elucidate the fundamental principles of complex biological systems amidst the ever-growing deluge of biological big data. As artificial intelligence (AI) continues to achieve disruptive breakthroughs in life sciences research, an AI-driven new paradigm is imminent. Through in-depth analysis of exemplary cases of AI-driven life sciences research, this article proposes the connotations and key elements of this new paradigm, and elaborates on and discusses the research frontiers under this new paradigm as well as the challenges confronting China.

Full Text

Preamble

Feature: Vigorously Promote the Transformation of Scientific Research Paradigms

Citation: Li X, Yu H C. A new paradigm of life science research driven by artificial intelligence. *Bulletin of Chinese Academy of Sciences*, 2024, 39(1): 50-58, doi: 10.16418/j.issn.1000-3045.20231211001.

Authors: Li Xin[1],[2], Yu Hanchao[3]*

[1]Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[2]ChinaXiv
[3]Bureau of Frontier Sciences and Education, Chinese Academy of Sciences, Beijing 100864, China

*Corresponding author

Abstract

The rapid development of biotechnology and information technology has ushered life sciences into a new era of data explosion, where traditional research paradigms struggle to reveal the fundamental principles of complex biological systems within ever-growing biological big data. As artificial intelligence (AI) continues to achieve disruptive breakthroughs in life science research, an AI-driven new paradigm is emerging. This article proposes the concept and key elements of this new paradigm through in-depth analysis of typical examples, elaborates on the research frontiers under this new paradigm, and discusses the challenges China faces.

Keywords: scientific research, life science, artificial intelligence, big data, scientific paradigm

---

1. Typical Examples of AI-Driven Life Science Research

Life is a complex system characterized by multiple hierarchical levels, multiple scales, dynamic interconnections, and mutual influences. Traditional life science research paradigms, when confronted with such extreme complexity, multi-scale spanning, and spatiotemporal dynamics, typically approach problems from a local perspective, establishing limited associations between biomolecules and phenotypes through experimental verification or restricted-level omics data analysis. However, even at enormous cost, these approaches usually discover only single linear associations under specific conditions, which differ significantly in complexity from the nonlinear nature of biological activities [1], making comprehensive understanding of entire networks difficult.

AI technologies, particularly deep learning and pre-trained large models, demonstrate superior pattern recognition and feature extraction capabilities. With vast parameter stacking, they can surpass human rational reasoning to better understand complex biological system patterns from data. Modern biotechnology continues to generate exponential growth in life science data, and the global research community has accumulated substantial experimentally described and validated data, creating a foundation for AI to decipher the underlying principles of life [2]. When equipped with sufficient high-quality data and life-science-appropriate algorithms, AI models can predict "high-dimensional" information and patterns from "low-dimensional" data across multiple hierarchical levels, achieving leaps from low-dimensional data like gene sequences and expression to high-dimensional complex biological processes. This enables the parsing of complex nonlinear relationships, such as the generation patterns of biological macromolecules, gene expression regulation mechanisms, and even complex biological systems involving multiple intersecting factors like individual development and aging. Under this trend, life science has recently witnessed several typical examples of AI-driven research, including protein structure prediction and gene regulation pattern analysis.

1.1 Protein Structure Prediction

Proteins, as key functional executors in organisms, have structures that directly affect vital biological processes including transport, catalysis, binding, and immune functions. While sequencing technology reveals amino acid sequences, any known protein chain could theoretically fold into an astronomical number of possible conformations, making accurate structure determination a long-standing challenge. Traditional techniques like nuclear magnetic resonance, X-ray crystallography, and cryo-electron microscopy require years to map a single protein's shape, are expensive and time-consuming, and offer no guarantee of success. Capturing the underlying principles of protein folding for accurate structure prediction has thus remained a major challenge in structural biology.

AlphaFold 2 employs attention mechanism-based deep learning algorithms trained on vast protein sequence and structure data, integrating prior knowledge from physics, chemistry, and biology to construct a protein structure prediction model comprising feature extraction, encoding, and decoding modules [3]. At the 2020 Critical Assessment of Protein Structure Prediction (CASP14), AlphaFold 2 achieved remarkable results, with protein 3D structure prediction accuracy comparable to experimental methods. This breakthrough brings three major transformations to life sciences.

First, it directly impacts drug discovery. Most drugs exert effects by binding to specific protein domains to alter protein function. AlphaFold 2 can rapidly calculate structures of numerous target proteins, enabling targeted drug design for effective binding [4].

Second, it enables rational protein design. Once AI deeply understands protein folding principles, this knowledge can design protein sequences that fold into desired structures. This allows biologists to freely design and modify proteins or enzymes according to needs, such as creating more active gene-editing enzymes [5] or even entirely novel protein structures not found in nature [6]. It also advances understanding of how genetic code information projects onto protein structures, greatly enhancing humanity's ability to engineer life.

Third, AlphaFold 2 fundamentally transforms the research paradigm in protein structure determination, shifting from time-consuming traditional experimental techniques to a new paradigm of low-barrier, high-precision, high-throughput prediction. This demonstrates that combining protein knowledge with AI can extract and learn high-dimensional, complex knowledge, promoting deeper understanding of protein physical structures and functions.

1.2 Gene Regulation Pattern Analysis

The Human Genome Project, one of the three major scientific projects of the 20th century, unveiled the prelude to life's mysteries. Although genetic information is stored in DNA sequences, each cell's fate and phenotype vary dramatically due to unique spatiotemporal contexts. This complex process is controlled

by sophisticated gene expression regulation systems, and exploring universal gene regulatory mechanisms represents one of the most important questions in life sciences following the genome project. Different cells' gene expression profiles provide an ideal window for understanding gene regulatory activities within biological systems. However, comprehensively interpreting gene regulation mechanisms through biological experiments alone requires capturing control experiments across different individuals, cell types, and environmental contexts. Traditional bioinformatics methods can only process small datasets and struggle to capture complex nonlinear relationships in large-scale, high-dimensional biological data lacking accurate annotations.

Recent breakthroughs in natural language processing, particularly large language models, offer new solutions by enabling models to understand human language-described knowledge through training data. Several international research teams have applied this approach to life sciences, constructing life foundation models capable of understanding gene dynamic relationships based on tens of millions of human single-cell transcriptome profiles and massive computational resources. These models—including GeneCompass [7], scGPT [8], Geneformer [9], and scFoundation [10]—use advanced algorithms like Transformer and diverse biological knowledge. Trained on underlying life activity information such as gene expression, these models learn to understand associations between "low-dimensional" life science data and complex "high-dimensional" mechanisms like gene regulatory networks and cell fate transitions, enabling effective simulation and prediction from low-dimensional data to high-dimensional information. This simulation of gene regulatory networks demonstrates excellent performance across broad downstream tasks, providing entirely new pathways for understanding gene regulation patterns.

These successful AI-driven life science research cases demonstrate that AI can break through traditional methodological limitations, construct theoretical projection systems from basic biological levels to entire life systems, and further advance life sciences to higher stages, heralding a new research paradigm.

2. Connotation and Key Elements of the New Paradigm

With continuous biotechnology advancement and rapid life science data growth, AI has demonstrated deep understanding and generalization capabilities for life science knowledge, elevating research scope and depth while propelling life science from the experiment-dominated first paradigm into an AI-driven new paradigm—the fifth paradigm (Figure 1 [Figure 1: see original paper]).

Through deep analysis of AI-driven life science research examples, we propose that the new paradigm resembles an intelligent new energy vehicle. Analogous to core automotive systems like battery, electronic control, motor, driver assistance, and chassis, the new paradigm requires five key elements: life science big data, intelligent algorithm models, computing platforms, expert prior knowledge, and interdisciplinary research teams (Figure 2 [Figure 2: see origi-

nal paper]). Just as the battery system provides energy, life science big data provides fundamental resources; algorithm models function like intelligent electronic control systems, enabling deep understanding of biological system mechanisms; computing platforms serve as the motor system, processing massive scientific data and complex computational tasks; expert prior knowledge acts as the driver assistance system, providing directional guidance and implementation experience; and interdisciplinary research teams function as the chassis, integrating diverse domain knowledge and skills through cross-disciplinary collaboration to enhance research efficiency and drive life science development.

2.1 Key Element 1: Life Science Big Data

Life science big data forms the "battery system" of the new paradigm vehicle. With new biotechnology development, life science big data characterized by multi-modality, multi-dimensionality, distributed storage, hidden associations, and multi-level integration has gradually emerged. Only through effective integration and innovative AI-driven data mining can we break through human cognitive limitations, generate new discoveries, and expand exploration boundaries. Examples include medical vision large models [11-13] that integrate multi-source, multi-modal, multi-task medical imaging data to enable various applications under few-shot and zero-shot conditions, and cross-species life foundation model GeneCompass [7], which effectively integrates global open-source single-cell data to achieve panoramic learning of gene expression regulation patterns across over 120 million single-cell training datasets.

2.2 Key Element 2: Intelligent Algorithm Models

Intelligent algorithm models constitute the "electronic control system" of the new paradigm vehicle. Extracting new patterns and knowledge from vast life science big data requires innovative AI algorithms and models. Developing life-science-appropriate AI algorithms, extracting effective biological features, and constructing large-scale dynamic biological process models represent central challenges [14]. Early examples include Gerstein team's application of Bayesian network algorithms to predict protein-protein interactions, published in *Science* and establishing a foundation for classical machine learning in bioinformatics [15]; graph convolutional neural networks analyzing protein-protein interaction networks [16] and gene regulatory networks [17]; and AlphaFold 2's use of Transformer models for rapid, high-accuracy protein structure calculation—all demonstrating the critical importance of AI algorithm models in the new paradigm.

2.3 Key Element 3: Computing Platforms

Computing platforms serve as the "motor system" of the new paradigm vehicle. Computing power forms the foundation for AI implementation. The continuous development of AI algorithm models suitable for life science research, such as deep learning and large models, demands increasingly powerful and efficient computing platforms. Future development should focus on building

hardware platforms supporting AI-enabled life science research, including high-speed, high-capacity storage systems, specialized chips for life science data processing, and dedicated processors for accelerating biological model inference and training. This will provide efficient, reliable computing capabilities to handle massive life science data and meet complex model construction requirements, ensuring AI application and innovation in life sciences.

2.4 Key Element 4: Expert Prior Knowledge

Expert prior knowledge functions as the "driver assistance system" of the new paradigm vehicle. In the new paradigm, existing life science knowledge provides valuable training constraints, important background information, and feature relationships for AI algorithms, helping interpret life science data complexity and validate AI applications. This knowledge plays crucial guiding roles in algorithm design and model construction, promoting more accurate and efficient solutions to life science problems. For instance, embedding life science expert prior knowledge and human-annotated information encoding in novel gene expression pre-training models [7] enhances interpretation of complex feature associations in biological data, demonstrating superior model performance.

2.5 Key Element 5: Interdisciplinary Research Teams

Interdisciplinary research teams form the "chassis system" of the new paradigm vehicle. Under the new paradigm, a multidisciplinary team comprising AI experts, data scientists, biologists, and medical scientists is essential for achieving breakthrough life science discoveries. Diverse backgrounds and close collaboration integrate specialized knowledge across AI, biology, and medicine, providing multiple perspectives and methods. This creates a solid foundation for comprehensively understanding complex mechanisms, offers more possibilities for innovative solutions, and drives breakthrough discoveries in life sciences.

3. New Paradigm-Enabled Frontiers and China's Challenges

Traditional research paradigms explore life like looking at a leopard through a tube, with biologists working in isolation across different subfields. As the new paradigm develops, life science research will embrace a novel modality characterized by AI prediction, guidance, hypothesis generation, and verification, giving rise to rapidly advancing frontier research directions.

3.1 Frontier Research Directions Enabled by the New Paradigm

Structural Biology. Current AI applications like AlphaFold remain at the "sequence-to-structure" stage for protein structure prediction and design [6,18,19], unable to simulate and predict protein structure and function under complex physiological conditions. The emergence of higher-quality, larger-scale

protein data and novel algorithms will enable systematic analysis of biomacromolecular structure and function across different physiological states and spatiotemporal conditions, achieving intelligent structure determination and precise design from "sequence-to-function" and even "sequence-to-multi-scale interactions."

Systems Biology. Current omics data analysis remains limited to lower-dimensional biological observations, without forming full-dimensional observations from gene level to cellular, individual, or even population levels. The new paradigm will integrate multi-dimensional, multi-modal biological big data with expert prior knowledge, extract key phenotypic features, construct multi-scale biological process models, and reconstruct underlying operational rules of complex biological systems, forming fundamental and broadly applicable new systems biology research paradigms.

Genetics. With multi-omics data accumulation and novel gene large models, genetics has entered a rapid development phase driven by the new paradigm. Self-supervised pre-training large models based on gene expression profile data will become powerful tools for parsing gene regulation patterns and predicting disease targets [9], expanding genetics research boundaries.

Drug Design and Development. With AlphaFold and molecular dynamics models, AI is already used for predicting and screening drug candidates. The new paradigm will further advance this field toward AI-assisted full-process drug design and development systems capable of autonomously optimizing drug structure and properties, simulating efficacy and safety, and generating efficient synthesis and production protocols, greatly accelerating drug development.

Precision Medicine. AI technologies including computer vision, natural language processing, and machine learning have permeated subfields like biological imaging, medical imaging, intelligent disease analysis, and target prediction. For example, AI diagnostic systems can already match or even exceed experienced clinicians in accuracy [20]. However, current models suffer from data bias, poor robustness, and low generalizability. New paradigm-driven universal precision medicine models will enable faster, more accurate disease diagnosis, molecular mechanism parsing, and therapeutic target discovery, improving human health.

3.2 Challenges Facing China's Life Science Research New Paradigm

Despite increased investment in life sciences, China faces significant challenges in establishing and promoting the new paradigm, including lack of high-quality life science data resource systems, insufficient AI key technologies and infrastructure, and scarcity of cross-disciplinary innovation ecosystems.

Lack of High-Quality Life Science Data Resource Systems. Chinese scientists still rely on foreign high-quality data in some frontier fields, while domestic data construction and utilization lag behind. China's life science data resources suffer from uneven distribution, requiring better coordination and

integration for efficient aggregation and systematic improvement. Additionally, data security, particularly biological data privacy and safety, urgently needs strengthening during collection, transmission, and storage.

Insufficient AI Key Technologies and Infrastructure. China lacks core technologies for the AI-driven new paradigm, with original algorithms, models, and tools requiring substantial development. Advanced computing and analysis methods for massive, high-dimensional, sparsely distributed life science big data are urgently needed. Future development should create hardware, software, and new computing media more suitable for life science applications, exploring novel computing-biology interaction modes. The new paradigm demands new comprehensive capabilities for data, networking, and computing, requiring accelerated development of next-generation information infrastructure to solve computing power bottlenecks.

Scarcity of Cross-Disciplinary Innovation Ecosystems. Current AI-driven life science research mostly follows a "small workshop" model of spontaneous group combinations, lacking the cross-disciplinary environment needed for new paradigm development. The U.S. National AI R&D Strategic Plan (2023 update) emphasizes the importance of interdisciplinary AI research. The new paradigm should encourage broader multi-disciplinary "grand crossing" and "grand integration," establishing novel research modes combining dry and wet lab work, theory and practice, while continuously cultivating high-level interdisciplinary research talent.

China has begun widely deploying and promoting interdisciplinary development, as highlighted in the *14th Five-Year Plan for National Economic and Social Development and Long-Range Objectives Through 2035*. Integrating AI-enabled life science paradigm transformation into China's national development vision will create comprehensive effects and establish more open new research ecosystems.

4. Conclusion

Life sciences are experiencing unprecedented transformation driven by biotechnology, information technology, and especially AI advancement. This shift moves from traditional hypothesis- and experiment-driven paradigms toward big data- and AI-driven research. This evolution will broadly transform scientific research activities across multiple levels, including epistemology, methodology, research organization, economics, and ethics.

We stand in an era of transformation and promise, where life science innovation and technological progress jointly map a future of deeper exploration into life's mysteries. With further development of general AI, life science research will soon achieve a new mode integrating dry and wet lab work with human-AI collaboration, ushering in a scientific era where AI autonomously abstracts new knowledge and patterns—"predicting what humans have not seen, thinking what humans have not thought."

References

1. Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 2023, 620: 47-60.
2. Erbe R, Gore J, Gemmill K, et al. The use of machine learning to discover regulatory networks controlling biological systems. *Molecular Cell*, 2022, 82(2): 260-273.
3. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 2021, 596: 583-589.
4. Borkakoti N, Thornton J M. AlphaFold2 protein structure prediction: Implications for drug discovery. *Current Opinion in Structural Biology*, 2023, 78: 102526.
5. Huang J Y, Lin Q P, Fei H Y, et al. Discovery of deaminase functions by structure-based protein clustering. *Cell*, 2023, 186(15): 3182-3195.e14.
6. Madani A, Krause B, Greene E R, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 2023, 41(8): 1099-1106.
7. Yang X D, Liu G L, Feng G H, et al. GeneCompass: Deciphering universal gene regulatory mechanisms with knowledge-informed cross-species foundation model. (2023-09-28). https://www.biorxiv.org/content/10.1101/2023.09.26.559542v1.
8. Cui H T, Wang C, Maan H, et al. scGPT: Towards building a foundation model for single-cell multi-omics using generative AI. (2023-05-01). https://www.biorxiv.org/content/10.1101/2023.04.30.538439v1.
9. Theodoris C V, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. *Nature*, 2023, 618: 616-624.
10. Cui H T, Wang C, Maan H, et al. scFoundation: A large-scale foundation model on single-cell transcriptomics. (2023-05-31). https://www.biorxiv.org/content/10.1101/2023.05.29.542705v1.
11. Moor M, Banerjee O, Abad Z S H, et al. Foundation models for generalist medical artificial intelligence. *Nature*, 2023, 616: 259-265.
12. Li C Y, Wong C, Zhang S, et al. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. (2023-06-02). https://arxiv.org/abs/2306.00890.
13. Zhang S, Xu Y, Li C Y, et al. BiomedGPT: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. (2023-10-01). https://www.biorxiv.org/content/10.1101/2023.09.27.559179v1.
14. Alber M, Buganza Tepole A, Cannon W R, et al. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *NPJ Digital Medicine*, 2019, 2: 115.
15. Jansen R, Yu H Y, Greenbaum D, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 2003, 302: 449-453.
16. Wu Y F, Gao M, Zeng M, et al. BridgeDPI: A novel Graph Neural Network for predicting drug–protein interactions. *Bioinformatics*, 2022, 38(9): 2571-2578.

17. Gan Y L, Hu X, Zou G B, et al. Inferring gene regulatory networks from single-cell transcriptomic data using bidirectional RNN. *Frontiers in Oncology*, 2022, 12: 899825.
18. Lutz I D, Wang S Z, Norn C, et al. Top-down design of protein architectures with reinforcement learning. *Science*, 2023, 380: 266-273.
19. Watson J L, Juergens D, Bennett N R, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 2023, 620: 1089-1100.
20. Kermany D S, Goldbaum M, Cai W, et al. Leveraging big data and AI in medical diagnosis. *Nature Medicine*, 2023, 29: 1723-1731.

Note: Figure translations are in progress. See original paper for figures.