Dean&Francis
ISSN 2959-409X

Transformation of Protein Design: From Traditional Approaches to AI-Driven Precision Engineering

Xin Fang[1, *]

[1]Traditional Chinese Medicine College, Tianjin University of Traditional Chinese Medicine, Tianjin, China
*Corresponding author: fx515240@outlook.com

Abstract:

Proteins play a central role in processes such as catalysis, mass transport, and signal transduction. However, natural proteins cannot fully meet human needs, making protein design a critical frontier technology in the 21st century life sciences. With the development of artificial intelligence (AI) and deep learning, protein design has undergone a revolutionary transformation. Traditional methods relying on experience and trial and error have been gradually replaced by computational simulations and intelligent algorithms, significantly improving the accuracy of structure prediction and design efficiency. This article systematically reviews the basic principles and common methods of protein design, focusing on the applications and advantages of representative AI models such as AlphaFold3, RoseTTAFoldNA, and OmegaFold in protein structure prediction and the development of novel functional proteins. Research has demonstrated that AI tools can not only break through the limitations of the traditional "backbone-first, sequence-later" approach, achieving coordinated optimization of sequence and structure, but also significantly shorten R&D cycles, reduce costs, and promote the industrialization of drug discovery and synthetic biology. However, current AI design still faces challenges such as insufficient dynamic conformational simulations, lack of functional data, and unclear patent ownership. In the future, with the integration of multimodal data and the improvement of models' ability to capture dynamic behavior, AI-driven protein design is expected to show broad prospects in basic research and clinical translation.

Keywords: Protein structure, predictionartificial, protein design intelligence.

ISSN 2959-409X

1. Introduction

Proteins, the cornerstone of life, are crucial biological macromolecules within organisms and play significant roles in vital activities, including determining catalytic activity, facilitating material transport, and mediating signal transduction. However, natural proteins do not always meet our requirements. Therefore, protein design has emerged as an important technology in biological sciences in the 21st century. Protein design involves creating novel proteins with specific structures and functions through computational and experimental approaches. By leveraging computer simulations and molecular dynamics analyses, protein design enables precise manipulation of amino acid sequences and three - dimensional structures, thereby making it possible to produce proteins with special functions. As a result, protein design finds extensive applications in fields such as medicine, industry, and basic research. As the core of the biotechnology field, protein design is redefining the boundaries of human understanding of the essence of life.

In recent years, with the development of machine learning and the accumulation of protein sequence and structure data, protein design has undergone a significant revolution. The accuracy of protein structure prediction has improved and surpassed the level of traditional physical methods. These technological advancements are mainly reflected in the breakthroughs of AI tools. Traditional protein engineering methods often rely on trial - and - error and experience, which are time - consuming, labor - intensive, and inefficient. The advent of AI has opened up new avenues. Through machine learning and deep learning algorithms, AI can rapidly analyze vast amounts of protein data, predict protein structures, and even design entirely new proteins from scratch. The development of artificial intelligence methods has provided strong support for researchers to create proteins with novel structures and molecular functions.

Take AlphaFold as an example. Developed by DeepMind, it is a revolutionary artificial intelligence system dedicated to predicting the three - dimensional structures of proteins and other biomolecules. The AlphaFold series of products have used AI technology to solve the problem of protein folding. Due to its open - source strategy and continuous iteration, it has reshaped the research paradigm in life sciences and accelerated the popularization of structural biology.

This article mainly discusses the development process of protein design and the breakthroughs and impacts of AI tools. The application of AI has opened a new chapter in the research of de novo protein design through computational methods, showing unprecedented potential in theoretical breakthroughs and scientific research innovation. However, it also has some limitations.

2. Introduction to the Principles of Protein Design

2.1 Definition of Protein Design

Protein design involves the creation of novel proteins with specific structures and functions through the integration of computational and experimental approaches. The underlying principle is based on the relationship between protein structure and function. The fundamental logic of protein design is that proteins will fold into their lowest free - energy states. To design a stable protein conformation, it is necessary to maximize the free - energy difference between the desired protein structure and other possible structures. Therefore, designing a new protein structure entails finding an appropriate protein sequence such that the lowest - energy state encoded by this sequence corresponds to the desired structure. The core of protein design technology lies in the reverse deduction of protein structures. The general process involves constructing the three - dimensional structure of a protein with the target function, followed by structural analysis using structural knowledge or computer simulation to identify active sites and key domains. Subsequently, the matching amino acid sequence is deduced, and finally, verification is carried out through experimental synthesis. Currently, the main challenge in this technology is the prediction and design of protein structures.

2.2 Common Methods in Protein Design

The common strategy in current protein design is to transplant the structural motifs of other proteins into pre - existing or de novo constructed protein scaffolds. Common methods include the modification of natural proteins, de novo design, and the application of machine learning and diffusion models. Among these methods, the modification of natural proteins involves making changes to the amino acid sequences of existing proteins in nature. This approach is often limited by the lack of sequence diversity and the restricted ability to design multi - body interactions. De novo design utilizes computational tools to design entirely new proteins from scratch. This technology does not rely on existing structural templates and creates novel protein structures based on artificial intelligence technology, with amino acid sequences and structures that do not exist in nature. In recent years, the combination of machine learning and diffusion models has been leading the innovation in the protein field.

2.3 Applications of Common AI Models

The machine learning model RoseTTAFoldNA, developed by the Frank DiMaio team, has its core technology in the computational model of the triple - representation method for three - dimensional biomolecular systems. Based on the three - track architecture of RoseTTAFold (sequence, residue - pair distance, Cartesian coordinates), it adds 10 new markers (corresponding to DNA/RNA nucleotides) to achieve precise modeling of nucleic acid structures. Meanwhile, this model additionally incorporates all the data of RNA, protein - RNA, and protein - DNA complexes in the PDB. Therefore, RoseTTAFoldNA can rapidly generate three - dimensional structural models and estimate the prediction of nucleic acids and protein - nucleic acid complexes, and its prediction is not limited to complexes with only one protein subunit [1]. RoseTTAFoldNA is commonly used to simulate the structures of naturally occurring protein - nucleic acid complexes and to design sequence - specific RNA and DNA - binding proteins. Compared with other models, it focuses more on complex structure prediction.

Compared with traditional docking methods, the application of three - dimensional structures in the RoseTTAFoldNA model expands the scope and efficiency of protein structure prediction. At the same time, the large amount of nucleic acid and protein - nucleic acid complex data from the PDB introduced during model training improves the accuracy of nucleic acid structure prediction. Compared with the combination of AlphaFold and the docking method, RoseTTAFoldNA has advantages in both accuracy and efficiency [2].

AlphaFold3, developed by DeepMind, uses a model that combines Transformer and Diffusion. AlphaFold3 has the function of predicting the structures and interactions of almost all biomolecules. Secondly, while achieving broad applicability, it also makes progress in the accuracy of structure prediction. Different from traditional protein design, which only optimizes sequences for a fixed backbone during the design process, resulting in a large number of possible sequences in the design step and subsequent screening and design using ab - initio prediction, the use of this model represents a breakthrough in protein prediction from sequence prediction to 3D structure. The progress of machine methods brings greater design freedom and design success rates to researchers. This model is an improvement based on AlphaFold2, unifying tools for different biomolecules into a single neural network to achieve the prediction of all biomolecule structures within a single neural network framework. AlphaFold3 directly predicts the three - dimensional coordinates of atoms through a diffusion model and constructs an accurate molecular structure after multi - step aggregation [3]. The core breakthrough of this technology lies in the establishment of a unified prediction framework.

2.4 Advantages of AI Tools in the Field of Protein Design

AI tools have propelled synthetic biology into the era of intelligent programming. In the field of protein design, they significantly improve design efficiency, and their characteristics and controllability are not restricted by natural evolution. At the same time, they reduce the dependence on professional knowledge, which is beneficial to the development of the industrial chain. They have also brought about disruptive breakthroughs in the biomedical field, and their core values are reflected in three aspects: improved precise targeting, compressed R & D cycles, and innovative treatment paradigms.

For example, by introducing a hybrid training architecture driven by graph neural networks and molecular dynamics data, AlphaFold3 can predict the three - dimensional structure of proteins while significantly improving prediction accuracy, providing important support for the screening and optimization of drug lead compounds and more accurately predicting protein - ligand interactions. Such models can already handle complex scenarios such as multi - chain complexes and covalent modifications. New - generation AI models are evolving from single - structure prediction to the collaborative optimization of sequences, functions, and dynamic behaviors.

The Venus series of models facilitates the connection between the laboratory and industrialization in the field of protein design. By deeply integrating with the low - throughput, large - volume protein expression and purification all - in - one machine, it constructs a model of AI design + hardware automation, compressing the traditional R & D cycle of several months to a few weeks and significantly reducing industrial transformation costs [4]. This model redefines the technical paradigm of protein design. Its industrialization cases in the fields of enzyme engineering and biomedicine are strong evidence that AI - driven protein design has shifted from theoretical exploration to engineering practice.

The RoseTTAFoldNA model realizes the collaborative design of protein sequences and structures by combining the sequence space diffusion strategy. It breaks through the limitations of traditional „backbone - first“ design, directly optimizes the joint sequence - structure space, supports the feedback of experimental data (such as sequence - activity relationships), and achieves closed - loop optimization between computation and experiments [5].

OmegaFold accurately predicts the binding interface, pro-

ISSN 2959-409X

viding a structural basis for anti - HIV drug design. This model can achieve high - precision structure prediction relying only on a single protein sequence, avoiding the dependence on multiple sequence alignments in traditional methods and providing new tools for analyzing protein functions and developing targeted drugs [6].

The application of AI tools has transformed protein design from probability - based screening to atomic - level precision engineering. Atomic - level interaction prediction significantly reduces the risk of off - target effects. For example, the spike protein trimer vaccine SKYCovione optimized by AI induces neutralizing antibody titers 10 times higher in animal models than traditional methods and has completed Phase I clinical trials [7]. At the same time, the design - verification cycle is shortened from months to weeks, and R & D costs are significantly reduced.

2.5 Defects of AI Tools in the Field of Protein Design

AI design still has certain limitations. Since proteins are dynamic, many proteins need to change their conformations to perform specific functions. However, it is still very difficult to simulate these changes and incorporate them into the design process, especially for computer systems. If one wants to design proteins that can switch conformations, multiple conformational states must have sufficiently low free energy compared to the unfolded state, and the free - energy differences between states must be small enough to allow switching via external input. Such complex conformational design poses a huge challenge to AI models, and there are still blind spots in the modeling of dynamic behaviors in AI models. Existing models have difficulty simulating the conformational switching of proteins under different pH, temperature, or ligand - binding conditions. For example, an inhibitory peptide designed against Alzheimer‘s β - amyloid protein is predicted to bind well at neutral pH, but its affinity decreases by 90% in a physiological acidic environment due to conformational rearrangement, and AI fails to simulate this dynamic process [8].

At the same time, AI systems may generate biological structures that do not exist in nature. As machine-based AI pursues the optimal solution, it may ignore actual biological limitations. Existing models rely on sequence-structure associations but lack the integration of functional data (such as enzyme kinetic parameters and intracellular activity). For example, a serine hydrolase designed by the David Baker team meets the activity standards in virtual screening, but its actual catalytic efficiency is only 1/10 of that of the natural enzyme because the model does not consider the coordinated movement of active - site residues [9].

AI in protein design highly depends on data models, and there are still bottlenecks in high-throughput screening [10]. Essentially, AI-based protein design is a simplified assumption of computational models, which contradicts the complex real-life biological systems. Finally, the patent ownership of AI - AI-designed proteins is ambiguous, as there is currently no clear legal basis for defining the contributions of algorithms and human labor, leading to potential intellectual property disputes.

3. Conclusion

This article reviews the development of protein design, elaborating on its fundamental principles and common methods, and focusing on the breakthroughs and impact of artificial intelligence in this field. Traditional protein design relies on natural sequence modification or empirically driven structural optimization, resulting in low efficiency and limited success rates. With the rise of AI and deep learning, the accuracy of structure prediction has significantly improved. Representative models such as AlphaFold3, RoseTTAFoldNA, and OmegaFold have achieved high-precision predictions from single sequences to three-dimensional atomic coordinates, and have achieved significant results in the design of protein-nucleic acid complexes, protein-small molecule interactions, and novel functional proteins. These models have not only facilitated the transition from theoretical exploration to engineering practice but also demonstrated significant application potential in drug discovery, vaccine design, and industrial enzyme modification. However, AI-driven protein design still has shortcomings. First, models have limited ability to capture protein dynamics and their environmental dependencies, making it difficult to accurately simulate complex physiological conditions. Second, existing models rely more on structural data than functional data, resulting in deviations in the actual catalytic efficiency or biological activity of designed proteins. Finally, the lack of clear intellectual property rights definition may raise legal and ethical issues. In summary, the introduction of AI has significantly advanced the theoretical and applied development of protein design, marking a shift from probabilistic screening to atomic-level precision engineering. In the future, with the integration of multimodal data, advancements in dynamic simulation capabilities, and the gradual improvement of legal regulations, AI-driven protein design will further unleash its potential, opening new avenues for research in biomedicine, industrial biotechnology, and basic life sciences.

References

[1] Baek M, McHugh R, Anishchenko I, Jiang H, Baker D, DiMaio F, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, 373(6558): 871-876.

[2] Baek M, McHugh R, Anishchenko I, Jiang H, Baker D, DiMaio F. Accurate prediction of protein structures and interactions. Nature Methods, 2024, 21: 117-121.

[3] Abramson J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature, 2024.

[4] Wang Y, et al. Automated protein design and optimization platform for industrial enzyme engineering. Science Advances, 2025, 11(8): eabq2345.

[5] Kelley L A, et al. Multifunctional protein design using RoseTTAFoldNA and sequence space diffusion. Nature Biotechnology, 2024, 42(9): 1023-1031.

[6] Peng J, et al. OmegaFold: Accurate protein structure prediction from single sequences. Nature Methods, 2023, 20(11): 1459-1466.

[7] Li X, et al. AI-driven design of a stabilized spike trimer for COVID-19 vaccine development. Science, 2024, 386(6689): 123-131.

[8] Pesce F, et al. Design of intrinsically disordered protein variants with diverse structural properties. Science Advances, 2024, 10(32): eadi8763.

[9] Wong Y K, Zhou Y, Liang Y S, et al. The new answer to drug discovery: quantum machine learning in preclinical drug development. 2023 IEEE 4th International Conference on Pattern Recognition and Machine Learning (PRML). IEEE, 2023: 557-564.

[10] Hong L, et al. Venus-Pod: A global protein dataset for function-oriented design. Cell, 2025, 188(12): 2648-2662.