Harnessing the Potential of Deep Learning to Improve Protein Structure Prediction: Challenges and Strategies

Yishuai Cheng

School of Medicine, Jiangnan University, Wuxi 214000, Jiangsu, China

Abstract. Proteins serve as the essential functional units of life, and understanding their three-dimensional structures is crucial for uncovering the biological mechanisms at play. Recent advancements in deep learning technologies, particularly those exemplified by AlphaFold, have transformed the conventional approaches to structure prediction. By utilizing a combination of residual neural network architectures and co-evolutionary features, these methods have achieved prediction accuracies that approach experimental results, ushering in a new era of intelligent modeling within computational biology. However, several core challenges persist in the realm of deep learning-driven structure prediction. These include the scarcity and quality issues surrounding training data, which can hinder model performance; the reliance on substantial computational resources, which limits the universality of algorithms; the difficulty in interpreting biophysical mechanisms due to the black-box nature of these models; and the challenges associated with sequence homology, particularly in accurately predicting the structures of orphan proteins. To address these issues, this paper proposes a range of multi-dimensional optimization strategies aimed at enhancing the efficacy of deep learning in protein structure prediction. In conclusion, the paper offers a forward-looking perspective on potential future research directions in this rapidly evolving field.

Keywords: Protein structure prediction; Deep learning; Challenges; Strategies

1. Background

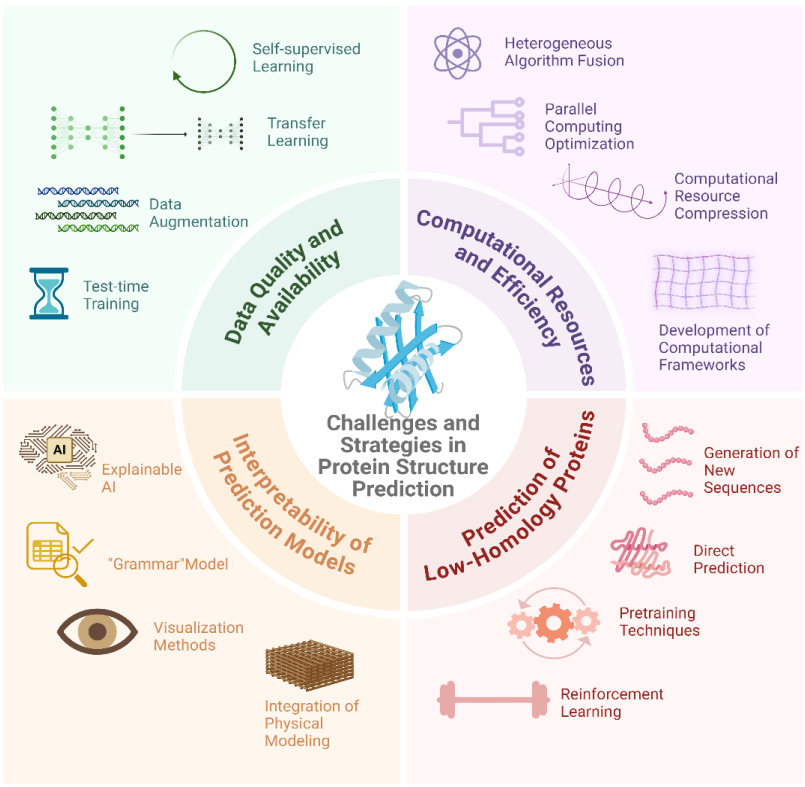

Figure 1 Challenges and Strategies for Protein Structure Prediction

Proteins are central molecules in biological processes, and their three-dimensional structures determine their biological functions. Resolving protein structures helps to understand their functional

mechanisms and is crucial for drug design and molecular biology research. Traditional protein structure prediction methods mainly include homology modeling and physics-based simulation. However, these methods have limitations: homology modeling is limited by the coverage of the known template structure database, and the accurate prediction rate for low-homology proteins decreases significantly [1-4]. Physics-based simulation methods have an extremely large computational magnitude and high computational costs in macromolecular systems [5]. These inherent limitations also promote the technological innovation in the field of computational biology.

In recent years, the breakthrough development of deep learning technology has provided new ideas for protein structure prediction. AlphaFold is a deep learning system developed by DeepMind. Through a deep residual neural network [6], it innovatively integrates multiple sequence alignment (MSA) and co-evolutionary features [7], achieving an average accuracy comparable to experimental structures [8]. Its core is to predict the distances and angles of protein groups in 3D space through a deep learning architecture [9]. In addition to AlphaFold, other deep learning methods are also emerging [10]. For example, DeepFold uses a template-free protein structure prediction based on a deep learning potential function [10]; AlphaFold2 uses sequence database search to construct MSA and then takes these features as inputs to a deep residual convolutional neural network [9]. This network architecture simplifies the training of deep networks by introducing shortcut connections with gating functions and utilizes the low-level inputs to high-level nodes in the network.

These breakthrough deep learning technologies are undoubtedly powerful computational tools in today's scientific research exploration. However, they still face challenges such as low data quality and availability, strict requirements for computational resources and efficiency, insufficient interpretability of prediction models, and difficulties in predicting low-homology proteins. This paper reviews the main challenges and countermeasures of deep learning in protein structure prediction.

2. Challenges and Countermeasures in Protein Structure Prediction

2.1. Optimization Strategies for Data Quality and Availability

Deep learning models for protein structure prediction and drug design rely heavily on large-scale, high-quality training data. However, the acquisition of experimental data depends on X-ray crystallography, nuclear magnetic resonance spectroscopy, cryo-electron microscopy, etc. [11], which is extremely difficult and costly. In addition, the amount of available protein structure data is still limited, and the Protein Data Bank contains only about 229 K macromolecular structures [12]. Therefore, it is crucial to develop alternative methods such as self-supervised learning, transfer learning, and the use of generated data.

2.1.1 Self-supervised Learning

Self-supervised learning techniques focus on training models using unlabeled data, such as protein sequences or structures, without the need for explicit supervision. This approach enables the model to extract meaningful features even when there is a scarcity of labeled data. For instance, contrastive representation learning trains a model capable of differentiating between various protein sequences or structures, which aids in developing stable and generalizable representations. Additionally, the "Protein sequence representations Learned Using Structural information" (PLUS) model has demonstrated strong performance across a range of downstream tasks, showcasing its effectiveness in the field [13].

2.1.2 Transfer Learning

Transfer learning is a technique that starts with pre-training a model on a large dataset and then fine-tunes it on a smaller dataset. This approach leverages the knowledge gained from the extensive data to enhance the model's performance on the limited data, ultimately aiming to boost the accuracy of predicting protein-ligand structures. For instance, one can pre-train a model using a vast array of generated docking conformations from tools like HelixDock [14]. Additionally, semi-supervised

learning methods, including self-training and consistency regularization, are employed to make the most of both labeled and unlabeled data in the context of protein sequence modeling [15, 16].

2.1.3 Data Generation and Data Augmentation

Researchers have investigated the use of generated data, including Structural Equation Models (SEM), to enhance experimental findings [17]. A notable example is the AlphaFold database, which boasts 200 million generated protein sequence structures. [11]. In a similar vein, the HelixDock Database offers a substantial collection of docking complexes created using physics-based docking tools [14]. These generated datasets serve as valuable resources for training models, thereby enhancing their generalization capabilities. Additionally, the quality of generated data—characterized by factors such as size and diversity—has become a key area of study. For instance, data augmentation techniques can create new protein sequences or structures by applying transformations like rotation or translation to existing datasets [18]. Furthermore, new ligand-protein complexes can be produced by docking ligands to various similar binding pockets found in the Protein Data Bank (PDB) [11]. In addition, developing more effective data augmentation strategies, such as strategies including adversarial perturbations and unsupervised clustering, can further enhance model generalization [19].

2.2. Optimization Strategies for Computational Resources and Efficiency

Protein structure prediction is a complex task involving the processing of high-dimensional data and the solution of optimization problems. Although AlphaFold relies on high-performance Graphics Processing Unit (GPU) clusters to achieve accurate predictions, the high cost limits the accessibility of research institutions. Developing efficient deep learning algorithms while maintaining prediction accuracy has become the key in this field [20, 21]. Although researchers have proposed various data compression techniques, such as methods combining operator fusion, tensor fusion, and hybrid parallel computing to reduce the computational cost of AlphaFold2 [20], multiple methods still need to be integrated in the case of insufficient computational resources.

2.2.1 Heterogeneous Algorithm Fusion

Some studies have improved efficiency through a multi-method fusion strategy. For example, a method that combines the functions of AlphaFold with all-atom symmetric docking simulation to predict the structures of complex symmetric protein assemblies [22]. In addition, Graph Neural Networks (GNNs) have been constructed with AlphaFold and hybrid architectures, significantly improving the efficiency of protein structure prediction in low-resolution cryo-electron microscopy maps [23].

2.2.2 Parallel Computing Optimization

One study proposed a method called APACE, which optimized AlphaFold2 to enable it to run on high-performance computing platforms on a large scale, reducing the protein structure resolution time from days to minutes [24]. Another study proposed a method called ParaFold, which separates the Central Processing Unit (CPU) and GPU parts to achieve large-scale structure prediction and improve GPU utilization. In addition, a new technology of Branch parallelism (BP) has a significant effect on improving the calculation speed [25].

2.3. Strategies for Improving Model Interpretability

The "black-box" nature of deep learning models leads to the lack of interpretability in protein structure prediction. Although current models provide highly accurate predictions, they often lack clear biological explanations. For example, AlphaFold can predict structures very accurately but cannot clearly reveal the underlying physical principles of protein folding, thus hindering in-depth understanding of the potential protein folding mechanisms. Therefore, it is very important to improve model transparency to help biologists understand the prediction mechanisms.

2.3.1 Attention Mechanism Analysis

Some studies have proposed attention mechanism analysis to reveal how the model learns structural features from protein sequences [26 - 29], gain an in-depth understanding of which amino acid sites have the greatest impact on chemical structure-biological activity [27], or which protein sequences are most relevant to zoonotic transmission. In addition, the attention mechanism can be used to identify key protein-protein interaction sites [30].

2.3.2 Physical Modeling and Biophysical Principles

Integrating physical modeling is also a strategy to improve interpretability. Ensuring that structure predictions follow known biophysical principles can be achieved by incorporating structural information into protein language models. The protein structure transformer (PST) [31] improves the self-attention mechanism of pre-trained language transformers by combining structural information with a structure extractor module, resulting in a parameter-efficient model that consistently outperforms state-of-the-art protein sequence-based models. In addition, protein language models can capture rich, interpretable protein biological representations, including binding sites, structural motifs, and functional domains [28]. By using sparse autoencoders to extract and analyze these interpretable features, a deeper understanding of how the model learns from protein sequences can be obtained.

2.4. Technologies for Predicting Low-homology Proteins

Deep learning models typically depend on multiple sequence alignments (MSA) for accurate predictions. However, when it comes to proteins that exhibit low homology or are classified as orphan proteins, the absence of alignment data can significantly hinder prediction accuracy. This issue is especially critical in the fields of structural biology and the exploration of novel proteins. Consequently, enhancing the predictive capabilities for low-homology proteins has emerged as a pressing challenge that needs to be addressed. In recent years, advancements in protein language models have opened up new avenues to tackle this problem. These models are capable of predicting protein structures, functions, and interaction networks by analyzing single sequences, thereby overcoming the limitations of traditional homology modeling methods that rely heavily on sequence similarity [32]. Single-sequence-based prediction methods, such as protein language models, have achieved accuracy close to experimental structures in predicting protein structures and have been used in fields such as drug discovery and synthetic biology [33]. However, predicting the three-dimensional protein structure from the amino acid sequence is a complex task and still requires a large amount of training data and computational resources [31]. These limitations restrict the universal application of this technology in protein research.

2.4.1 New Sequence Generation

One method is to use deep learning models to generate new protein sequences and improve MSA. For example, using a transformer-based language model (MSA-Augmenter) to generate new sequence models and integrating protein-specific attention mechanisms and large-scale MSA to generate new protein sequences, retaining the co-evolutionary information from poor MSA, thus improving the quality of protein structure prediction [34].

2.4.2 Direct Prediction by Deep Learning

Another method is to use deep learning models to directly predict protein structures from the original MSA without manual management or sequence alignment [35]. The rawMSA model uses embeddings to adaptively map discrete inputs in the symbol dictionary to vectors in the continuous space, allowing it to capture complex evolutionary patterns in MSA. This method has demonstrated considerable advancements in predicting secondary structure, relative solvent accessibility, and residue-residue contact maps.

2.4.3 Generative Pretraining Technology

Researchers have also suggested the application of generative pretraining technology to enhance protein structure prediction [36]. The MSAGPT model employs a two-dimensional evolutionary position encoding scheme that effectively simulates the intricate evolutionary patterns found in multiple sequence alignments (MSA), resulting in a notable increase in the accuracy of structure predictions. Furthermore, incorporating tertiary structure information into MSA has been shown to enhance alignment accuracy, particularly for homologous sequences that are distantly related [37]. The MAFFT-DASH model takes this a step further by utilizing a combination of hybrid protein sequence and structure similarity scores, which significantly boosts MSA accuracy and surpasses the performance of other MSA methods in terms of precision.

2.4.4 Reinforcement Learning Optimization

To tackle these challenges, researchers have introduced hybrid approaches that merge physics-based simulations with deep learning techniques, particularly reinforcement learning optimization strategies [38]. These innovative methods enhance prediction accuracy and have found applications across diverse fields, including protein design and engineering. Future research directions should focus on broadening the use of protein language models for function prediction, particularly in areas like protein complex structures and interactions between proteins and small molecules [39]. This expansion has the potential to significantly enhance applications in biomedicine and drug discovery [40].

3. Future Prospects

Deep learning technology has significantly improved the accuracy of protein structure prediction through innovative architectures, breaking through the inherent limitations of traditional homology modeling and physics-based simulation. However, this field still faces multiple challenges: data scarcity promotes the development of self-supervised learning and generative pretraining technologies; the bottleneck of computational efficiency promotes the integration of sparse matrices and heterogeneous algorithms; the insufficient interpretability of the model gives rise to the parsing of attention mechanisms and the integration of physical constraints; the prediction of low-homology proteins is limited by the coverage of multiple sequence alignment (MSA) and needs to rely on single-sequence analysis of protein language models to make breakthroughs. From these perspectives, future research should focus on three dimensions: developing a cross-scale modeling framework that couples generative pretraining and reinforcement learning to achieve multi-level predictions from monomers to complexes; constructing an experiment-computation closed-loop verification system to dynamically optimize the model through data such as cryo-electron microscopy; exploring quantum-classical hybrid algorithms to break through the computational barriers of ultra-large-scale systems. Such progress will promote the transformation of deep learning from structure prediction to mechanism analysis, laying a solid foundation for the explosive development of biological gene technology.

References

[1] Mihăşan, M. and A.I. Cuza, *Basic protein structure prediction for the biologist: A review.* Archives of Biological Sciences, 2010. **62**: p. 857-871.

[2] Chang, Y., et al., *A Guide to In Silico Drug Design.* Pharmaceutics, 2022. **15**(1).

[3] Meier, A. and J. Söding, *Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling.* PLoS Comput Biol, 2015. **11**(10): p. e1004343.

[4] Kairys, V., M.K. Gilson, and M.X. Fernandes, *Using protein homology models for structure-based studies: approaches to model refinement.* ScientificWorldJournal, 2006. **6**: p. 1542-54.

[5] Vallat, B., C. Madrid-Aliste, and A. Fiser, *Modularity of Protein Folds as a Tool for Template-Free Modeling of Structures.* PLoS Comput Biol, 2015. **11**(8): p. e1004419.

[6] Wuyun, Q., et al., *Recent Progress of Protein Tertiary Structure Prediction.* Molecules, 2024. **29**.

[7] Ouyang-Zhang, J., et al. *Predicting a Protein's Stability under a Million Mutations*. in *Neural Information Processing Systems*. 2023.

[8] Yousef, M. and J. Allmer, *Deep learning in bioinformatics.* Turk J Biol, 2023. **47**(6): p. 366-382.

[9] Auslander, N., A.B. Gussow, and E.V. Koonin, *Incorporating Machine Learning into Established Bioinformatics Frameworks.* International Journal of Molecular Sciences, 2021. **22**.

[10] Pearce, R., et al., *Fast and accurate Ab Initio Protein structure prediction using deep learning potentials.* PLoS Computational Biology, 2022. **18**.

[11] Zhang, Z., et al. *Geometric Deep Learning for Structure-Based Drug Design: A Survey*. 2023.

[12] Berman H M, Burley S K. *Protein Data Bank (PDB): Fifty-three years young and having a transformative impact on science and society.* Quarterly reviews of biophysics, 2025, **58**: e9.

[13] Min, S., et al., *Pre-Training of Deep Bidirectional Protein Sequence Representations With Structural Information.* IEEE Access, 2019. **9**: p. 123912-123926.

[14] Liu, L., et al., *Pre-Training on Large-Scale Generated Docking Conformations with HelixDock to Unlock the Potential of Protein-ligand Structure Prediction Models.* ArXiv, 2023. **abs/2310.13913**.

[15] Golyadkin, M., et al., *Refining the ONCE Benchmark with Hyperparameter Tuning.* IEEE Access, 2023. **12**: p. 3805-3814.

[16] Cai, D., Z. Cai, and M. Li, *Self-supervised Reflective Learning through Self-distillation and Online Clustering for Speaker Representation Learning.* ArXiv, 2024. **abs/2401.01473**.

[17] Burley, S.K., et al., *RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning.* Nucleic Acids Research, 2022. **51**: p. D488 - D508.

[18] King, J.E. and D.R. Koes, *SidechainNet: An all‐atom protein structure dataset for machine learning.* Proteins: Structure, 2020. **89**: p. 1489 - 1496.

[19] Khan, A., et al., *A Survey of the Self Supervised Learning Mechanisms for Vision Transformers.* ArXiv, 2024. **abs/2408.17059**.

[20] Wang, G., et al., *HelixFold: An Efficient Implementation of AlphaFold2 using PaddlePaddle.* ArXiv, 2022. **abs/2207.05477**.

[21] Hu, B., et al. *Advances of Deep Learning in Protein Science: A Comprehensive Survey*. 2024.

[22] Li, Z., J. Yu, and Q. Ye, *SGNet: Folding Symmetrical Protein Complex with Deep Learning.* ArXiv, 2024. **abs/2403.04395**.

[23] Ma, X. and D. Si, *Beyond Current Boundaries: Integrating Deep Learning and AlphaFold for Enhanced Protein Structure Prediction from Low-Resolution Cryo-EM Maps.* ArXiv, 2024. **abs/2410.23321**.

[24] Park, H.S., et al., *APACE: AlphaFold2 and advanced computing as a service for accelerated discovery in biophysics.* Proceedings of the National Academy of Sciences of the United States of America, 2023. **121**.

[25] Zhong, B., et al., *ParaFold: Paralleling AlphaFold for Large-Scale Predictions.* International Conference on High Performance Computing in Asia-Pacific Region Workshops, 2021.

[26] Vig, J., et al., *BERTology Meets Biology: Interpreting Attention in Protein Language Models.* bioRxiv, 2020.

[27] Xu, S. and L. Xie, *Dumpling GNN: Hybrid GNN Enables Better ADC Payload Activity Prediction Based on Chemical Structure.* ArXiv, 2024. **abs/2410.05278**.

[28] Simon, E. and J.Y. Zou, *InterPLM: Discovering Interpretable Features in Protein Language Models via Sparse Autoencoders.* bioRxiv, 2024.

[29] Zhang, M. and Y. Cui, *Self-supervised learning-based emotion recognition using physiological signals.* Frontiers in Human Neuroscience, 2024. **18**.

[30] Wadhawan, K., et al., *Towards Interpreting Zoonotic Potential of Betacoronavirus Sequences with Attention.* ArXiv, 2021. **abs/2108.08077**.

[31] Chen, D., et al., *Endowing Protein Language Models with Structural Knowledge.* ArXiv, 2024. **abs/2401.14819**.

[32] Kyro, G.W., T. Qiu, and V.S. Batista. *A Model-Centric Review of Deep Learning for Protein Design*. 2025.

[33] Zheng, Y., et al., *Large Language Models in Drug Discovery and Development: From Disease Mechanisms to Clinical Trials.* ArXiv, 2024. **abs/2409.04481**.

[34] Zhang, L., et al., *Enhancing the Protein Tertiary Structure Prediction by Multiple Sequence Alignment Generation.* ArXiv, 2023. **abs/2306.01824**.

[35] Mirabello, C. and B. Wallner, *rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments.* PLoS ONE, 2019. **14**.

[36] Chen, B., et al., *MSAGPT: Neural Prompting Protein Structure Prediction via MSA Generative Pre-Training.* bioRxiv, 2024.

[37] Rozewicki, J., et al., *MAFFT-DASH: integrated protein sequence and structural alignment.* Nucleic Acids Research, 2019. **47**: p. W5 - W10.

[38] Son, A., et al., *Integrating Computational Design and Experimental Approaches for Next-Generation Biologics.* Biomolecules, 2024. **14**.

[39] Desai, D., et al., *Review of AlphaFold 3: Transformative Advances in Drug Design and Therapeutics.* Cureus, 2024. **16**.

[40] Wei, J., et al., *Protein–RNA interaction prediction with deep learning: structure matters.* Briefings in Bioinformatics, 2021. **23**.