Dean&Francis

ISSN 2959-409X

Advances in Protein Structure Prediction through Deep Learning Models

Xinyang **Zhang** [1*]

[1] *School of Chemistry and Molecular Biosciences, The University of Queensland, St Lucia, Brisbane QLD 4072, Australia*

*Corresponding author: xinyang.zhang2@student.uq.edu.au

Abstract:

Accurate protein structure prediction (PSP) is essential for understanding biological function. However, experimental determination of protein structures remains costly and limited in scope. In recent years, advances in deep learning (DL) have substantially improved PSP. These methods help close the gap between the vast number of known protein sequences and limited experimentally resolved structures. This article reviews five representative DL architectures: convolutional neural networks (CNNs), recurrent neural networks (RNNs), transformers, graph neural networks (GNNs), and diffusion-based models. Key protein databases used for training and validation are also introduced, including the Protein Data Bank (PDB), UniProt, Pfam, RefSeq, and the Big Fantastic Database (BFD). Performance evaluations based on the Critical Assessment of Protein Structure Prediction (CASP) benchmarks show that models such as AlphaFold2 and its successors achieve near-experimental accuracy. Nevertheless, challenges remain for low-homology sequences, protein–protein interactions, and dynamic folding pathways. Current limitations of DL in PSP also include data bias, restricted interpretability, and an inability to fully capture protein dynamics. To overcome these barriers, future directions may include explainable AI, debiased datasets, and integration with molecular dynamics simulations.

Keywords: Protein Structure Prediction; Deep Learning; AlphaFold; Neural Network Architectures; CASP

1 Introduction

Accurate prediction of protein three-dimensional (3D) structures is crucial for understanding their biological functions. Misfolded proteins or improper interactions can disrupt cellular processes, contributing to a range of diseases and immune disfunctions [1]. Certain proteins are central to pathogen recognition, therefore serving as key targets in therapeutic drug discovery [1].

Current techniques such as x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy provide high resolution protein structures. However, these methods are often expensive, time-consuming and limited to a small subset of proteins, typically those well folded and stable proteins [2]. As a result, a substantial gap remains between the vast number of known protein sequences and known protein structures. As of mid-2025, The UniProt TrEMBL database contains over 252 million protein sequences [3], while the Protein Data Bank holds only approximately 239,000 resolved structures [4].

The protein folding problem has been recognized as three interconnected challenges: (1) deciphering the thermodynamic principles by which amino acid sequence determines its native structure (the folding code), a concept grounded in Anfinsen's dogma; (2) predicting the native 3D structure from sequence (structure prediction); (3) understanding how proteins fold so rapidly despite the vast number of possible conformations(folding kinetics), a kinetic challenge known as Levinthal's paradox [5].

These complexities have motivated the development of computational methods aimed at addressing these challenges. Methods can be categorized into template-based modelling (TBM), template-free modelling (TFM), and ab initio [1]. TBM involves constructing 3D models by aligning the target sequence to resolved homologous protein structures [1]. However, its performance heavily replies on the availability of suitable templates within structure database, limiting its application for novel sequences. This limitation prompted the development of template-free modelling (TFM). Unlike approaches that are reliant on predefined templates, TFM uses threading approach or fragment assembly techniques to build models [1]. However, although termed "template-free", most models of TFM are still trained on Protein Data Bank (PDB) datasets [1]. Lastly, ab initio methods are also called as free modelling because it can directly predict a protein's 3D structures from amino acid sequences [1]. This follows Anfinsen's thermodynamic hypothesis: the native conformation is the global minimum of free energy [6]. Despite its great advantages in terms of theory, it demands extensive computational resources because of vast conformational search space [1].

DL has been a revolution in machine learning for protein structure prediction (PSP). Methods such as AlphaFold2 or RoseTTAFold released in 2021 that maps amino acid sequence directly to 3D structure [2].

These methods employed multiple sequence alignments (MSAs) and Evoformer-based modules to predict with near-experimental accuracy [2]. With these advances, AlphaFold3 uses a diffusion-based framework for wider variety of biomolecular complexes [7]. ESMFold introduced pre-trained language models (PLMs) that embed evolutionary constraints in model parameters directly not dependent on MSAs anymore [2].

This paper introduced deep-learning-based approaches for PSP, including discussion of five representative neural network structures (Section 2), major databases for PSP (Section 3), evaluation measures implemented for validating model prediction (Section 4), along with concluding section on key challenges and future research directions in PSP (Section 5).

2 Deep learning models for PSP

Recent advances in DL have enabled the development of increasingly accurate models for protein structure prediction by integrating diverse data sources. Unlike traditional machine learning approaches that rely heavily on co-evolutionary signals from homologous sequences, DL models are able to learning complex structural features directly from its sequence. DL models typically process either MSAs or single sequences encoded via PLMs to predict protein structures using neural networks (Figure 1). This approach underpins the architecture of leading models such as AlphaFold2, which use attention-based networks and leverage concepts from large language models (LLMs) to achieve near-experimental accuracy. In the following sections, key DL model architectures for protein structure prediction were analysed.

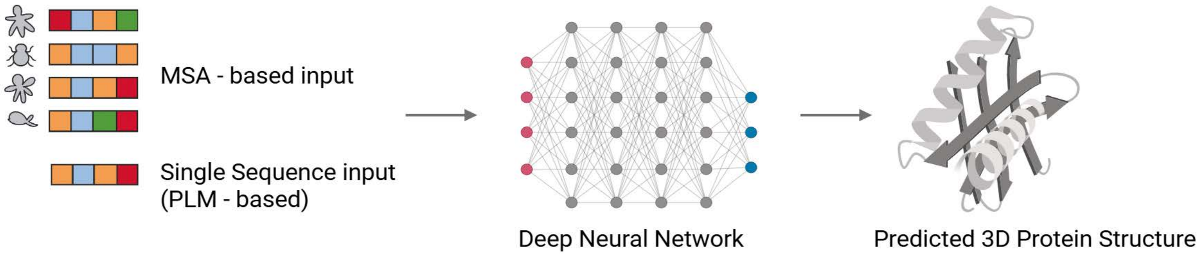

Fig. 1. General workflow of deep learning–based PSP. (Protein sequences embeddings PLMs are used as inputs to deep neural networks, These models learn spatial and structural relationships to generate 3D protein structures.) (Picture credit: Original)

ISSN 2959-409X

2.1 Convolutional neural networks

CNNs are characterized by convolutional operations that extract spatial features from input data [1]. Originally, CNN was developed for image recognition by efficiently capturing local spatial patterns through hierarchical layers of convolutions. This property enables CNNs to learn local spatial patterns in amino acid sequences. CNNs process amino acid sequences converted into 2D matrices using strategies like principal component analysis (Figure 2(A)) [1]. Successive convolution and pooling layers are applied to reduce noise, enhance relevant features, and abstract high-level structural patterns [1].

One successful application of CNNs in PSP was ProALIGN, which utilizes deep CNNs to enhance sequence–template alignments [8]. This model integrates both sequence and structural context improve alignment quality. It directly learns alignment likelihoods between residues from pairwise features such as PSSMs and inter-residue distances [8]. This significantly enhances the accuracy of template-based modeling, especially for distantly related proteins. Similarly, RaptorX utilizes deep ResNet structures for predicting inter-residue distances and orientations using MSA [9]. These models have shown promise in precise folding structure prediction and significant improvement in performance of TFM approaches [9].

2.2 Recurrent neural networks

RNNs are critical for processing sequential information. At each point in time, output depends on previous states [1]. In PSP, RNNs can capture sequential dependencies among amino acids as well as the evolutionary directions, making them more suited for modeling biologically ordered sequences (Figure 2(B)) [1]. However, traditional RNNs suffer several problems such as vanishing gradient problems and struggle with long-range dependencies.

To overcome these challenges, Long Short-Term Memory (LSTM) were developed [1]. It composed of cell blocks, each containing a memory cell. These memory cells allow the network to selectively retain or discard information (Figure 2(C)) [1]. As a result, LSTMs largely resolve this issue and improve the applicability of RNN-based models in the field of PSP.

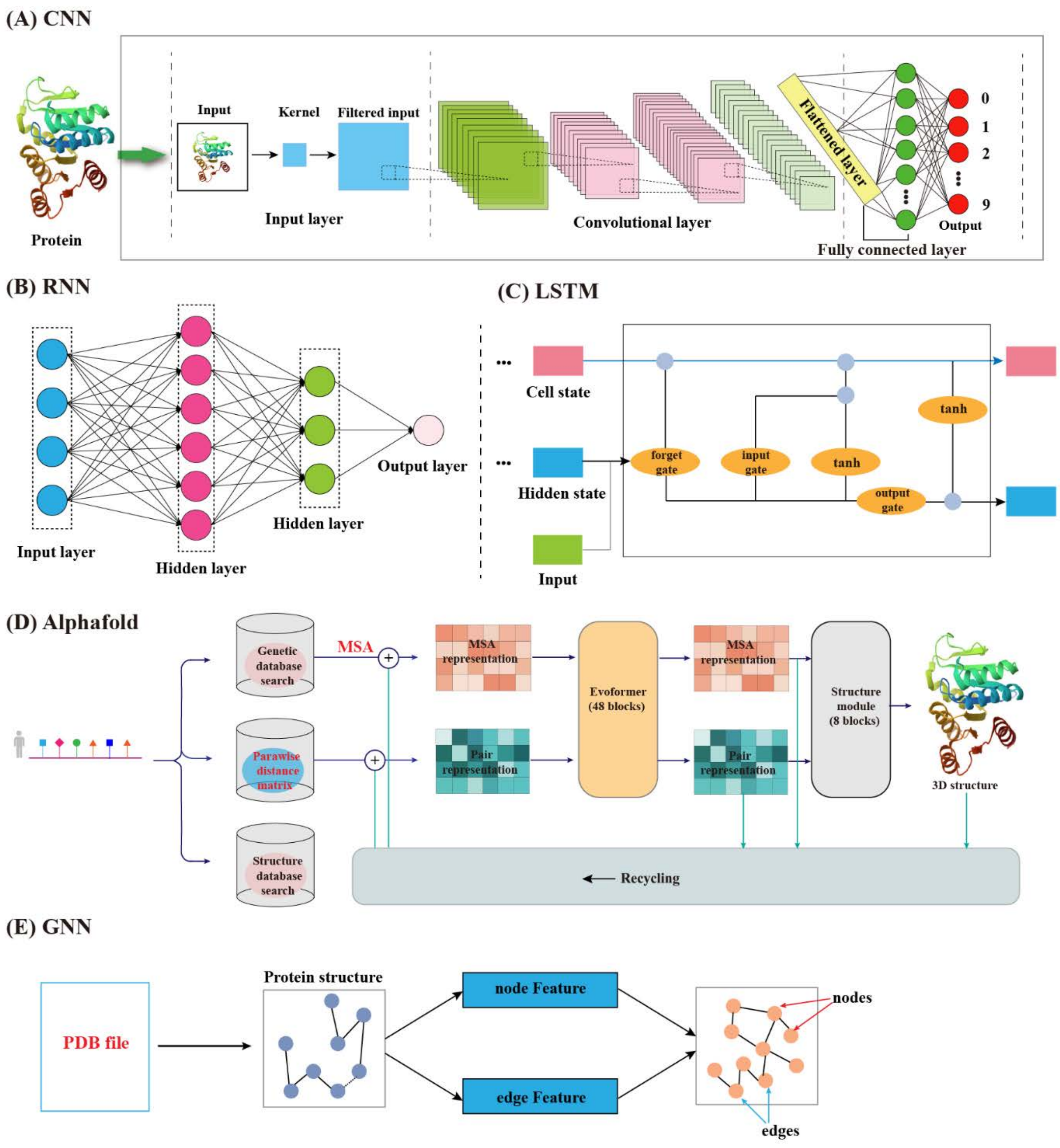

Fig. 2. Representative deep learning architectures applied to protein structure prediction. ((A) CNNs capture local spatial patterns from sequence-derived features through convolution layers, followed by fully connected layers for prediction. (B) RNNs model sequential dependencies in amino acid chains by propagating information across hidden states. (C) LSTM networks extend RNNs with memory cells and gating mechanisms to alleviate vanishing gradient issues and capture long-range dependencies. (D) AlphaFold integrates MSAs, pair representations, and the Evoformer module with iterative recycling to predict three-dimensional structures at near-experimental accuracy. (E) GNNs represent proteins as graphs, residues as nodes and inter-residue interactions as edges. This enables the analysis of both local and global structural features.) (Picture credit: Original)

In practice, researchers developed a bidirectional LSTM (BLSTM) model (termed CSI-LSTM) to predict threestate protein secondary structures using NMR shift data [10]. BLSTMs process forward and reverse sequence information, enabling them to capture broader context compared to unidirectional models [10]. Using a dataset of over 2,500 proteins, CSI-LSTM achieved a Q3 accuracy of 87.1%, outperforming earlier chemical shift–based methods such as CSI 2.0 and CSI 3.0 [10]. Meanwhile, researchers proposed a LSTM and BLSTM-based model to predict protein tertiary structures directly from primary amino acid sequences, specifically predicting amino acids occurrence using the MSA [11].

However, with rapid development of new architectures like transformers, RNN-based methods are gradually replaced due to slower training speeds and dependence on experimental inputs like NMR chemical shift.

2.3 Transformer

The transformer has been widely and effectively applied in language modelling tasks. This DL architecture replaces recurrent layer with self-attentions to capture long-range dependencies [12].

One groundbreaking example is AlphaFold. The model adopts the Evoformer module to handle relationships between protein sequences and structures [13]. Specifically, Evoformer contains multi-head self-attention, triangular updates, and pairwise residue features to iteratively learn residue spatial relations [13]. Besides, the model uses MSA) to extract co-evolutionary information from homologous sequences (Figure 2(D)) [13]. Nevertheless, AlphaFold2 heavily relies on deep MSAs for proteins with high evolutionary information [13].

To alleviate this, AlphaFold3 makes architectural modification. It replaces the Evoformer with the Pairformer module and applies geometric diffusion to sample complex conformations [7]. AlphaFold3 extends its application to more types of biomolecular assembly such as proteins, nucleic acids and chemically modified residues [7].

Another type of transformer-based architecture ESMFold trains large-scale transformer PLMs (ESM-2) to directly predict accurate structures from a single sequence without requiring MSAs [14]. This is especially beneficial when it comes to predicting proteins with few or no homologous sequences. Similarly, OmegaFold implements transformer modules combined with geometry-inspired modules for predicting 3D protein structure from single sequence [15]. Outcomes demonstrated strong performance even without evolutionary sequence alignment information available for guidance [15].

2.4 Graph Neural Networks (GNN)

GNN has received significant attention in diverse fields including natural language processing and bioinformatics (image) [1]. In PSP, the geometric configuration of proteins can be modeled as graph, where amino acid residues are nodes and interactions between residues are edges (Figure 2(E)) [1]. Compared with traditional methods that are solely based on amino acid sequences, GNN can capture spatial geometry of proteins more effectively for PSP. For example, researchers proposed GNN architecture based on ccPDB 2.0 dataset for secondary structure prediction [16]. The graph is constructed from primary sequence where edges representing residue interactions. They indicated that adopting graph representation in place of sequence representation enhanced the model's ability to learn structural features [16].

A more recent study introduced GraphGPSM model, which extend DL application in protein structure evaluation [17]. GraphGPSM adopts an equivariant graph neural network (EGNN) architecture that encodes both local residue-level features and global backbone features directly into the nodes and edges [17]. Through message-passing operations, node representations are iteratively refined during information exchange while preserving spatial equivariance during information propagation [17]. Compared with traditional scoring models such as REF2015 and lDDT-based approaches, GraphGPSM achieves superior performance in estimating global model accuracy [17]. Notably, the model also enhances multi-domain proteins structures predicted by AlphaFold2. Accurate prediction of these proteins typically depends heavily on MSAs and coevolutionary data, yet GraphGPSM demonstrated advantages in such challenging scenarios [17].

2.5 Diffusion – based model

RoseTTAFold introduced a deep learning architecture integrating MSAs templates, and geometric constraints to improve protein folding accuracy [18]. However, like AlphaFold2, the ability of RoseTTAFold in prediction is constrained when lacking evolutionary information [18]. To overcome this limitation, Rfdiffusion takes the existing RoseTTAFold network and extend it into a generative model using diffusion-based methods [19]. RFdiffusion iteratively transforms residue frames into realistic protein backbones [19]. This allows the model to generalize beyond PDB, creating protein folds and functional scaffolds with experimentally validated success. This model is capable of protein design across a wide range of tasks, such as monomer generation, binder design and symmetric oligomer construction [19]. Notably, RFdiffusion has been shown to outperform previous deep learning methods such as hallucination and joint inpainting [19].

While RFdiffusion focuses primarily on de novo protein design, EigenFold adapts diffusion models to PSP tasks [20]. EigenFold yields diverse structure ensembles instead of a single fold, capturing heterogeneity better to reflect protein dynamics [20]. The model introduces a harmonic diffusion process to model proteins as an ensemble of harmonic oscillators, iteratively improving conformations by cascading along the eigenmodes [20].

3 Protein Databases

Reliable protein databases form the foundation of computational PSP. They supply curated sequences, evolutionary classifications, structural references, and large-scale datasets, supporting model training, benchmarking, and validation.

The Protein Data Bank (PDB) is the central collections for experimentally determined 3D protein structures. These structures are primarily resolved with NMR spectroscopy and X-ray crystallography. Methods like AlphaFold2 greatly depend on PDB structures for data training and assessing in CASP competitions.

The Universal Protein Knowledgebase (UniProtKB) is critical for protein sequence and functional annotation [3]. It integrates ontology-based resources such as Gene Ontology (GO), Rhea (biochemical reactions), and ChEBI (chemical entities) to enhance functional annotation [3]. Machine learning applications such as ProtNLM have also been developed to enhance prediction capacity. For PSP, UniProt is the primary sequence source for models such as AlphaFold and ESMFold [7, 13, 14].

Pfam classifies proteins into families and domains using profile hidden Markov models (HMMs) [21]. Its latest release covers most UniProt sequences. In PSP, Pfam offers domain-level annotations which enhances model training and function assessment purpose [21].

RefSeq provides a curated non-redundant genomic DNA, RNA and protein database that collects genomes of thousands of species across all domains [22]. The main aim is to provide reference standards for genes. To achieve this, RefSeq integrates multiple approaches, including automated annotation pipelines and manual expert curation [22]. It relies on multiple sources for annotation which includes computational analysis and collaboration with researchers' community [22]. RefSeq records are assigned unique accession identifiers (e.g., NM, NP, XP) and annotated with gene names, functional features, and domain-level information [22].

The Big Fantastic Database (BFD) built from metagenomic and environmental datasets, comprises billions of clustered entries which detects remote homologs and generate deep MSAs for model trainings, such as AlphaFold2 and AlphaFold3 [7, 13].

4 Model Validation

The Critical Assessment of protein Structure Prediction (CASP) has been widely used for evaluating PSP, capturing both global fold similarities and local atomic precision. This uses blind experiments to compare predicted structures against experimental resolved structures. Metrics used for evaluation includes Global Distance Test–Total Score (GDT-TS), Template Modeling Score (TM-score), Root Mean Square Deviation (RMSD), and local residue accuracy measures such as plDDT/lDDT [23]. Additionally, performance also depends on MSA depth, computational efficiency, and scalability across different protein classes [23].

GDT-TS quantifies the percentage of residues within 1–8 Å of the experimental structure [23]. CASP XV reported median GDT-TS values approaching 90% for the majority of monomeric targets, meaning many predicted structures were almost identical in accuracy experimental resolution [23]. Nonetheless, the accuracy dropped sharply for proteins with shallow sequence alignments (<0.1 alignment depth-to-length ratio) [23]. For example, the aphid effector protein T1131 achieved a best GDT-TS below 40%, demonstrating the limitations of low-homology sequences [23].

TM-score measures global structural similarity without being affected by protein length, ranging from 0 (random) to 1 (perfect) [23]. CASP XV showed AlphaFold2 pipelines achieved TM-scores above 0.9 for most monomeric targets, particularly when enhanced sampling strategies like deeper MSAs and extended recycling applied [23]. By contrast, default implementations such as ColabFold often underperformed, emphasizing the necessity of tuning for high-accuracy predictions [23]. Additionally, ResNet-based models in RaptorX achieved average TM-scores of approximately 0.64–0.67 on CASP13 free-modelling targets, outperforming traditional ab initio methods [10]. Nevertheless, transformer-based systems such as AlphaFold2 outperformed earlier CNN-derived approaches, with GDT-TS of around 90% at CASP14, equivalent to TM-scores above 0.9 [23].

RMSD measures the average deviation between backbone atom positions in predicted and experimental structures [23]. In CASP 15, over 90% of models achieved Cα RMSD values below 3 Å, and nearly 40% fell below 1 Å, underscoring the near-experimental accuracy of current models [23]. AlphaFold2-derived architectures consistently outperformed all competitors, reaching sub-angstrom (<1 Å) accuracy for stable monomeric proteins [13, 23]. However, their RMSD remained sensitive to flexible domain arrangements.

lDDT measures local accuracy of a predicted structure by comparing interatomic distances, without requiring superposition [23]. This makes it suitable for flexible or multi-domain proteins. AlphaFold2 introduced plDDT as per residue accuracy estimate. CASP XV demonstrated that plDDT values closely mirrored true lDDT scores in monomeric proteins, with mean deviations below 0.1 for core residues [23]. However, accuracy estimates were less reliable at interfaces, emphasizing the ongoing challenges in modelling protein–protein interactions [23]. At CASP14, AlphaFold2 demonstrated that pLDDT strongly correlated with true residue-level accuracy and proved highly effective for ranking alternative models, thereby enhancing prediction reliability [13].

5 Challenges and Future Directions

DL methods like AlphaFold2 have transformed protein structure prediction, achieving accuracy comparable to experimental methods. Nevertheless, several key challenges and limitations constrain their broader applicability in PSP.

One major challenge is the lack of interpretability in deep neural networks. Although confidence metrics such as predicted lDDT (pLDDT) provide residue-level accuracy estimates, the internal decision-making of transformer architectures remains opaque. This limits the biological interpretability of predictions. Future work may focus on explainable AI frameworks can trace predictions back to specific sequence features or evolutionary couplings [1].

A second challenge arises from data bias in training datasets. Current models are heavily reliant on structures deposited in PDB, which are dominated by soluble globular proteins and static conformations. This restricted coverage can lead to overfitting and weak generalization [1]. Consequently, predictions performance is worse for underrepresented categories such as membrane proteins, intrinsically disordered proteins, and proteins from non-model species. Future directions to reduce this bias include diversifying training datasets beyond the PDB by incorporating structures from synthetic datasets or debiased structural datasets.

Meanwhile, existing models also face limitations in capturing folding processes. Most predictors generate static structures while cannot address conformational dynamics and kinetic pathways. This gap is critical because many proteins function through conformational change or transient states that cannot be described by a single static model. Future integration of deep learning with molecular dynamics simulations, or simplified lattice frameworks such as Monte Carlo simulations and the hydrophobic–polar (HP) model could simulate folding thermodynamics at large scales and capture dynamic landscape of folding.

6 Conclusion

Recent deep learning breakthrough have transformed protein structure prediction, helping to close the gap between vast protein data and limited experimentally resolved structures. Architectures such as CNNs, RNNs, transformers, GNNs, and diffusion-based models have all contributed advantages to PSP. Complementary resources such as UniProt, PDB, and Pfam provide support for training, benchmarking, and validating these models. Progress has been evaluated through CASP assessments CASP. In these assessments, blind testings are used to measure PSP performance compared to experimental determined structures. Models such as AlphaFold2 and its successors have achieved near-experimental accuracy across many targets. However, limitations of low sequence homology and dynamic conformational changes still require further innovations. Although current methods remain constrained by interpretability, dataset bias, and the inability to capture folding pathways, explainable AI, debiased databases and integration with dynamic models would offer promising solutions.

References

1. Y. Meng, Z. Zhang, C. Zhou, et al, Protein structure prediction via deep learning: an in-depth review. Front. Pharmacol. 16, 1498662 (2025). https://doi.org/10.3389/fphar.2025.1498662

2. Z. Zhang, C. Ou, Y. Cho, et al, Artificial intelligence methods for protein folding and design. Curr. Opin. Struct. Biol. 93, 103066 (2025). https://doi.org/10.1016/j.sbi.2025.103066

3. The UniProt Consortium, A. Bateman, M-J. Martin, et al, UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. 53, D609–D617 (2025). https://doi.org/10.1093/nar/gkae1010

4. RCSB Protein Data Bank (RPD Bank), PDB Statistics: Overall Growth of Released Structures Per Year. https://www.rcsb.org/stats/growth/growth-released-structures (accessed 30 Jul 2025).

5. K.A. Dill, S.B. Ozkan, T.R. Weikl, et al, The protein folding problem: when will it be solved? Curr. Opin. Struct. Biol. 17, 342–346 (2007). https://doi.org/10.1016/j.sbi.2007.06.001

6. C.B. Anfinsen, Principles that govern the folding of protein chains. Science. 181, 223–230 (1973). https://doi.org/10.1126/science.181.4096.223

7. J. Abramson, J. Adler, J. Dunger, et al, Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 630, 493–500 (2024). https://doi.org/10.1038/s41586-024-07487-w

8. L. Kong, F. Ju, W-M. Zheng, et al, ProALIGN: Directly learning alignments for protein structure prediction via exploiting context-specific alignment motifs (2020).

9. J. Xu, M. McPartlon, J. Li, Improved protein structure prediction by deep learning irrespective of co-evolution information. Nat. Mach. Intell. 3, 601–609 (2021). https://doi.org/10.1038/s42256-021-00348-5

10. Z. Miao, Q. Wang, X. Xiao, et al, CSI-LSTM: a web server to predict protein secondary structure using bidirectional long short term memory and NMR chemical shifts. J. Biomol. NMR. 75, 393–400 (2021) https://doi.org/10.1007/s10858-021-00383-9

11. J. Antony, A. Penikalapati, J.V.K. Reddy, et al, Towards protein tertiary structure prediction using LSTM/BLSTM, in Advances in Computing and Network Communications, eds. S.M. Thampi, E. Gelenbe, M. Atiquzzaman, et al, Springer, Singapore, pp. 65–77 (2021).

12. A. Vaswani, N. Shazeer, N. Parmar, et al, Attention is all you need (2017).

13. J. Jumper, R. Evans, A. Pritzel, et al, Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583–589 (2021) https://doi.org/10.1038/s41586-021-03819-2

14. Z. Lin, H. Akin, R. Rao, et al, Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 379, 1123–1130 (2023) https://doi.org/10.1126/science.ade2574

15. R. Wu, F. Ding, R. Wang, et al, High-resolution de novo structure prediction from primary sequence (2022).

16. T.H. Nahid, F.A. Jui, P.C. Shill, Protein secondary structure prediction using graph neural network, in Proceedings of the 2021 5th International Conference on Electrical Information and Communication Technology (EICT), IEEE, Khulna, Bangladesh, October 2021, pp. 1–6 (2021)

17. G. He, J. Liu, D. Liu, G. Zhang, GraphGPSM: a global scoring model for protein structure using graph neural networks. Brief. Bioinform. 24, bbad219 (2023) https://doi.org/10.1093/bib/bbad219

18. M. Baek, F. DiMaio, I. Anishchenko, et al, Accurate prediction of protein structures and interactions using a three-track neural network. Science. 373, 871–876 (2021) https://doi.org/10.1126/science.abj8754

19. J.L. Watson, D. Juergens, N.R. Bennett, et al, De novo design of protein structure and function with RFdiffusion. Nature. 620, 1089–1100 (2023) https://doi.org/10.1038/s41586-023-06415-8

20. B. Jing, E. Erives, P. Pao-Huang, et al, EigenFold: Generative protein structure prediction with diffusion models (2023)

21. J. Mistry, S. Chuguransky, L. Williams, et al, Pfam: The protein families database in 2021. Nucleic Acids Res. 49, D412–D419 (2021) https://doi.org/10.1093/nar/gkaa913

22. T. Goldfarb, V.K. Kodali, S. Pujar, et al, NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Res. 53, D243–D257 (2025) https://doi.org/10.1093/nar/gkae1038

23. A. Kryshtafovych, T. Schwede, M. Topf, et al, Critical assessment of methods of protein structure prediction (CASP)—Round XV. Proteins. 91, 1539–1549 (2023) https://doi.org/10.1002/prot.26617