Machine Learning Methods for Protein Structure Prediction: A Systematic Literature Review

Hassan Tariq*, Areej Fatima, Muhammad Sohaib

Department of Computer Science
University of Agriculture Faisalabad Pakistan
E-mails: hassan@uaf.edu.pk, areejfatima7644@gmail.com, sohaibafzal102@gmail.com

**Corresponding author*

Received: June 03, 2025 **Accepted: November 26, 2025**

Published: June 30, 2026

Abstract: *Protein structure prediction (PSP) is a fundamental challenge in computational biology, essential for understanding molecular mechanisms and accelerating drug discovery. This systematic review, conducted under the PRISMA guidelines, presents the application of machine learning (ML) methods for PSP, with a focus on deep learning models and hybrid approaches from 2014 to 2025. A comprehensive search across major databases, retrieved 1,939 studies, of which 43 met the inclusion criteria for full-text analysis. The studies reviewed employed state-of-the-art ML techniques such as Convolutional Neural Networks (CNNs), Support Vector Machines (SVMs), Random Forests (RF), and ensemble models. Advanced methods like AlphaFold and RoseTTAFold were highlighted for their accuracy in tertiary structure prediction, with TM-scores surpassing 0.7. Other models like ThreaderAI and DeepMSA2 demonstrated significant advancements in template-based modeling and secondary structure prediction. The analysis identified common challenges, including dataset biases primarily linked to well-characterized proteins from the Protein Data Bank (PDB), limited performance in predicting intrinsically disordered proteins (IDPs), and the lack of interpretability in deep learning models. Few studies integrated Explainable AI (XAI) techniques to enhance model transparency, indicating an area for future development. In conclusion, this systematic review provides insights into current ML-driven methodologies for PSP, outlines key challenges, and suggests the need for improved dataset diversity, explainable models, and hybrid approaches to bridge the gap between prediction and biological interpretation.*

Keywords: *Protein structure prediction, Machine learning, AlphaFold, Deep learning, Computational biology.*

Introduction

Proteins are fundamental biological macromolecules responsible for various cellular processes, including enzymatic reactions, signal transduction, immune responses, and molecular transport. Their functional roles are directly linked to their three-dimensional structures, which are determined by their amino acid sequences, as described by Anfinsen's dogma [2]. The process of accurately predicting protein structures from sequences remains a key challenge in bioinformatics and structural biology, affecting drug discovery, disease modeling, and synthetic biology [4, 20].

Traditional experimental techniques for protein structure determination, such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy (cryo-EM), provide high-resolution insights but are limited by high costs, time consumption, and difficulties in handling large or flexible proteins [6]. Consequently, computational methods, particularly ML and DL approaches, have emerged as transformative tools in PSP.

In recent years, deep learning models such as AlphaFold and RoseTTAFold have redefined the

capabilities of computational PSP, achieving near-experimental precision for tertiary structure prediction [4, 20]. These models leverage evolutionary data, multiple sequence alignments (MSA), and structural templates to predict 3D conformations with high confidence. In particular, AlphaFold achieved a median global distance test (GDT) score of 92.4 on the CASP14 targets, setting a new benchmark for predictive accuracy [20].

In addition, innovations in protein language models (PLMs), such as ProtTrans and ESMFold (https://esmatlas.com/resources?action=fold), have introduced the concept of treating protein sequences as natural language, allowing predictions without the need for MSA [33]. These advances paved the way for single-sequence-based predictions, significantly reducing computational overhead and extending PSP capabilities to orphan proteins without ligationary alignment [13].

Despite these breakthroughs, challenges remain in modeling IDPs, improving the prediction accuracy of multidomain proteins, and addressing dataset biases introduced by reliance on well-characterized proteins from the PDB [6]. Addressing these limitations is critical for broadening the applicability of ML-based PSP models and enhancing their reliability for therapeutic applications.

This systematic review aims to evaluate advances in ML methodologies for protein structure prediction, with a focus on their applications, limitations, and future directions. Explore deep learning architectures, hybrid models, and single-sequence prediction techniques, offering a comprehensive analysis of their performance across various datasets, including CASP [32], PDB [34], and UniProt [39]. Furthermore, this review emphasizes emerging trends and unresolved challenges, providing insights into the next steps toward achieving more accurate and scalable PSP models.

Materials and methods

This study is supervised and reported under the outlined guidelines by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (www.prisma-statement.org).

Sources of information and search strategy

Articles and reports are retrieved from electronic databases including PubMed, Google Scholar, Nature, IEEE Xplore, ScienceDirect, and Semantic Scholar. The search targeted publications in English from 2014 to 2025. The query descriptors used were: “proteins”[MeSH Terms] AND “structure”[All Fields] AND “prediction”[All Fields] AND “machine learning”[MeSH Terms] AND “methods”[All Fields]. Filters ensured the search was restricted to articles with these terms appearing in the title.

Study selection criteria

The inclusion and exclusion criteria for selecting studies were as follows:

Inclusion criteria

- Studies published between 2014 and 2025.
- Articles focusing on machine learning methods for protein structure prediction.
- Full-text availability in peer-reviewed journals.
- Studies related to machine learning, deep learning, or neural network methods are included.

Exclusion criteria

- Conference papers, book chapters, and review articles were excluded.
- Studies not published in English were excluded.
- Articles not focused on machine learning methods for PSP were excluded
- Review articles were excluded unless they present novel meta-analyses or systematic insights.
- Paid articles were excluded.

A total of 1,939 articles were initially retrieved. After eliminating duplicates and applying inclusion criteria, 43 studies were selected for detailed analysis.

Selection of studies and data extraction

Two researchers independently selected studies by reviewing the titles and abstracts of the articles. Comprehensive evaluations of articles and case reports are conducted. Only the information relevant and useful in machine learning, deep learning, and neural network models for protein structure prediction is extracted and documented.

Data analysis

Key findings related to the methodologies, machine learning models, and outcomes relevant to protein structure prediction are properly organized and interpreted. To provide a comprehensive understanding of current advancements and future directions, critical insights like model performance, applicability, and trends are documented properly.

ML techniques analysis

The review primarily focused on state-of-the-art ML techniques employed for protein structure prediction, including:

- CNNs: for capturing spatial hierarchies in protein data.
- AlphaFold and RoseTTAFold: achieving near-experimental accuracy in tertiary structure prediction.
- PLMs: such as ProtTrans and ESMFold, which predict structures without MSAs.
- RNNs and LSTM models: for sequence-based learning.
- Hybrid models: Combining multiple ML methods to enhance prediction accuracy.

Evaluation metrics

The performance of ML models was evaluated using several metrics:

- TM-Score: Measures the structural similarity between predicted and native structures.
- RMSD: Quantifies the average distance between atoms of superimposed proteins.
- pLDDT: Used in AlphaFold to estimate the confidence in prediction.
- GDT: Evaluates the overall accuracy of the 3D models.

These metrics provided a benchmark for comparing the effectiveness of different ML architectures in predicting primary, secondary, and tertiary protein structures.

Results

A comprehensive search in multiple databases within the past 10 years produced a total of 1,939 results: PubMed (287 hits), Google Scholar (1,060 hits), ScienceDirect (186 hits), Nature (93 hits), IEEE Xplore (29 hits), and Semantic Scholar (360 hits). After the elimination of duplications, 1,406 unique results were retrieved. Based on titles and abstracts, 500 relevant articles were identified and selected. 376 were excluded due to access restrictions (paid articles), and 81 were excluded as they were literature reviews or conference papers. At last, 43 studies were included for full-text analysis.

These studies spanned diverse areas of protein structure prediction, including primary, secondary, tertiary, and quaternary structures, as well as specific applications like binding site and loop predictions. The studies employed cutting-edge machine learning techniques such as neural networks, SVMs, and ensemble models. Databases such as PDB [34], CASP benchmarks [32], and UniProt [39] were frequently used for model training and validation. Significant outcomes include improvements in prediction accuracy, computational efficiency, and the identification of novel protein features that were previously difficult to model.

AlphaFold [32] and RoseTTAFold [5] set new standards with TM-scores beyond 0.7 in tertiary structure prediction, showcasing the progress in computational methods. Single-sequence prediction methods and innovations in PLMs highlighted a significant shift from traditional multisequence alignment techniques, reflecting the growing potential of deep learning and AI in protein structure prediction.

These studies provide a useful perspective on the evolving role of AI in the advancement of protein structure research. The integration of machine learning with biological data promises to open new avenues for drug discovery, protein engineering, and understanding disease mechanisms at a molecular level.

So, inclusion criteria led to the selection of 43 publications that were read in full (Fig. 1).

The Table 1 provides a comprehensive summary of the studies included in this review, offering detailed insights into the contributions of each study to the field of protein structure prediction using machine learning and neural network algorithms. Each study was selected based on its focus on protein structure and its innovative application of computational methods to improve accuracy and efficiency in protein analysis. The Table 1 outlines key information, including study authors, publication year, specific protein structure or characteristic analyzed, machine learning techniques used, datasets used, and significant results. This overview highlights the diversity of approaches, from traditional machine learning methods such as SVMs and random forests to more advanced deep learning techniques such as CNNs and RNNs, underscoring their effectiveness in addressing complex biological problems.

Discussion

This systematic review provides a comprehensive overview of the recent advances in ML and DL models for PSP. Over the past decade, the field has undergone a remarkable transformation, moving from template-based comparative modeling to powerful AI-driven predictions that can achieve near-experimental accuracy. The widespread application of neural networks, particularly CNNs, RNNs, and transformer architectures, has not only accelerated the prediction pipeline but has also significantly improved precision, especially in tertiary structure prediction.

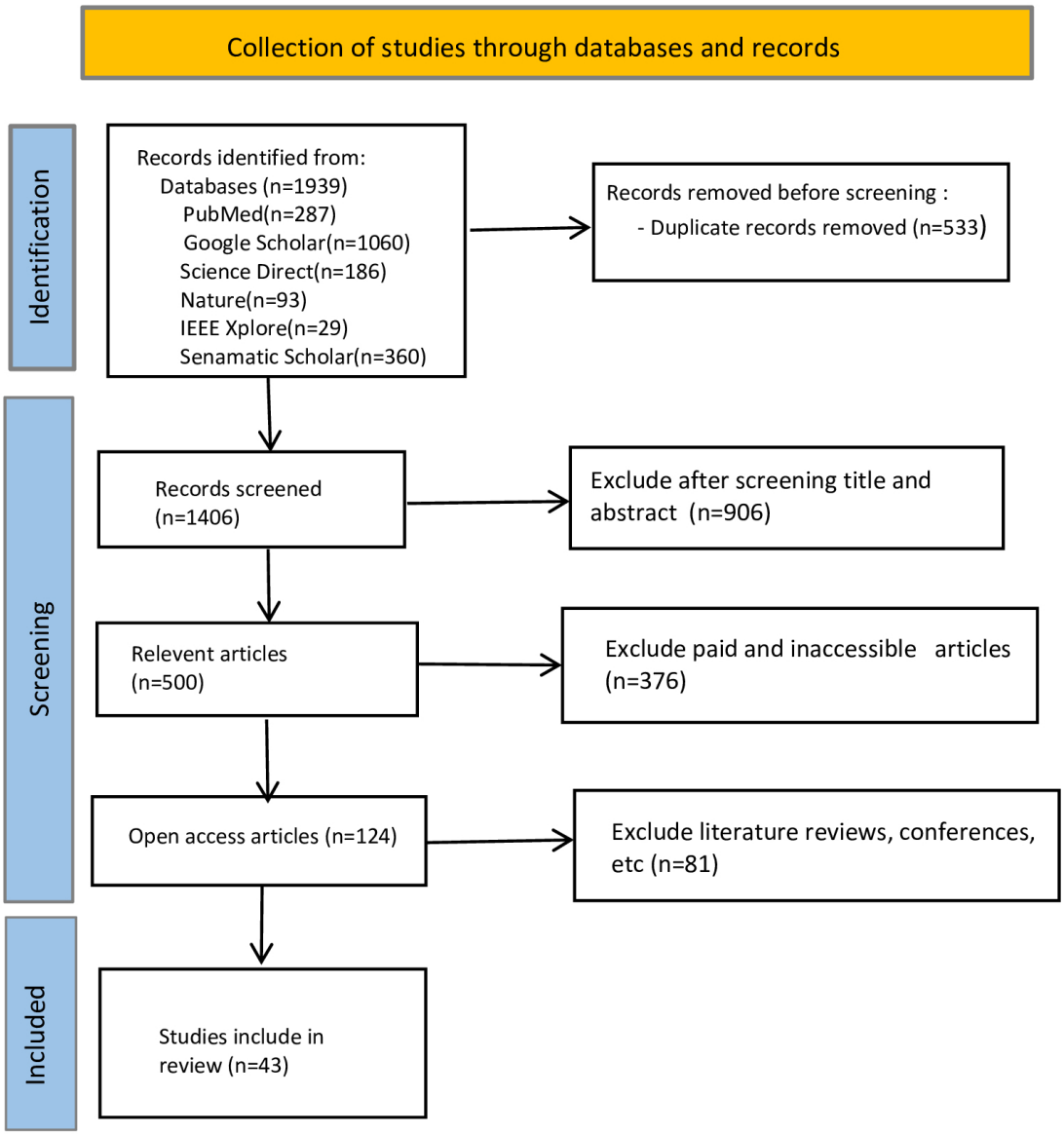

Fig. 1 PRISMA workflow diagram

The most significant breakthrough was achieved with the release of AlphaFold2 [4], which revolutionized PSP by introducing a three-track neural network that integrates sequence, MSA, and structural information. AlphaFold2 demonstrated a median GDT score of 92.4 in CASP14, pushing the boundaries of computational biology. Similarly, RoseTTAFold [20] followed a similar architecture, using attention mechanisms and MSA embeddings to produce highly accurate models. These tools mark a transition in the field from alignment-based heuristics to deep learning-based end-to-end predictions.

Another critical evolution is the emergence of PLMs, such as ProtTrans [11], ESM-1b and ESM-Fold [26], and HelixFold-Single [13]. These models treat amino acid sequences as language tokens and utilize transformer architectures trained on millions of proteins to infer structural patterns. PLMs enable single-sequence-based prediction, eliminating the need for MSAs. This is particularly impactful for orphan proteins, which lack homologs and cannot benefit from MSA-driven inference. PLM-based approaches like RaptorX-Single and RGN2 have begun to show promising results in such cases [6, 33].

Table 1. Summary of each study included in systematic review

Sr#	Study	Objective	ML method	Dataset	Key results
1	[1]	Improve secondary structure prediction through ensemble learning	Ensemble of SVM, Naive Bayes, and Decision Tree	RS126	Achieved 91.4% accuracy using majority voting; ensemble outperformed individual models
2	[3]	Investigate deep learning methods for protein secondary structure prediction (PSSP) using primary sequences	CNNs, BiLSTM, multi-scale CNNs	CullPDB, CB513	Achieved Q8 accuracy of 70.4%, slightly exceeding the previous best result of 70.3%
3	[5]	Advance the field of structural biology by improving prediction techniques using modern DL models	AlphaFold, AlphaFold-Multimer, Geometric Deep Learning, AF2Complex, OmegaFold, ESMFold, Transfer Learning, Monte Carlo Tree Search	AF2Complex	Reported improvements in plDDT scores and median benchmark scores across multiple protein targets
4	[6]	Emphasize the importance of open-access data (PDB) in enabling AI-based PSP research	No specific model used	PDB	PDB plays a crucial role in AI integration and structural biology advancements
5	[7]	Review and highlight DL advancements for PSP in bioinformatics	CNN, Stacked Sparse Autoencoder, Conditional Neural Fields, Deep Belief Networks	CullPDB, CB513, CASP9, CASP10, RS126, 25PDB	Achieved Q3 accuracy: 82.34% (RS126), 84.32% (CB513), 83.76% (25PDB); 74.2% SOV score on CASP9/10 datasets
6	[8]	Improve PSSP accuracy using Hessian Free Optimization with simple FFNNs	Hessian Free Optimization, SVM	PISCES, CASP13	Achieved 80.4% Q3 and 0.77 SOV on PISCES; 78.14% Q3 and 0.755 SOV on CASP13.
7	[9]	Develop a high-performing PSP system without relying on MSAs	AminoBERT	UniParc, SCOPe v1.75, Uniclust30, De novo proteins	Outperformed AlphaFold2 and RoseTTAFold on orphan proteins

Sr#	Study	Objective	ML method	Dataset	Key results
8	[10]	PISCES, CASP13	Develop SERTStructNet for improved secondary structure prediction	SERTStructNet (D-CNN, SENet, BiGRU, BiLSTM, Transformer)	Achieved 84.9% Q3 accuracy and 85.1% SOV score; outperformed RaptorX-SS and JPRED
9	[12]	Apply CNN and ML classifiers for secondary structure prediction	CNN, SVM, Random Forest, Naive Bayes	RS126, CB513, 25PDB	CNN-SVM achieved Q3: 82.34% (RS126), 84.32% (CB513), 83.76% (25PDB); improved accuracy by up to 2.83%
10	[13]	Develop HelixFold-Single to predict PSP from primary sequences without MSAs using PLMs	HelixFold-Single, PLM, Masked Language Model	UniRef30, PDB, Uniclust30, AlphaFold DB, CAMEO	Achieved competitive TM-scores; demonstrated effectiveness of MSA-free structure prediction
11	[14]	Improve 8-state PSSP accuracy using ensemble ML approach	SVM, kNN, Decision Tree, Naive Bayes, Random Forest, PCA, LDA	CB6133, CB513	RF with boosting and PCA achieved 66.52% accuracy and 0.85 AUC; DT showed highest precision (0.7119)
12	[15]	Propose DeepACLSTM for accurate 8-state secondary structure prediction	DeepACLSTM (ACNN + BLSTM)	CB513, CASP10, CASP11, CB5534	Achieved 75.0% Q8 accuracy on CASP10 and 73.0% on CASP11; outperformed baseline models
13	[16]	Combine AlphaFold with physics-based molecular dynamics refinement to improve PSP	AlphaFold + Molecular Dynamics	CASP13	Refined AlphaFold models improved TBM scores (61.7 vs. 44.6) and FM-scores (69.0 vs. 67.2) over CASP13 benchmarks
14	[17]	Develop a cascaded CNN-LSTM model for effective PSSP feature extraction	CNN, LSTM	CB513, CB6133, CASP10, CASP11, PISCES and Scratch	Q8 scores: 73.35 (CB513), 75.17 (CB6133), 75.09 (CASP10), 72.80 (CASP11), 73.84 (PISCES), 76.47 (Scratch)
15	[18]	Develop a single-sequence PSP method using PLMs and structure generation module	RaptorX-Single (ESM-1b, ESM-1v, ProtTrans + DL architecture)	PDB, SAbDab, IgFold-Ab, SAbDab-Ab, Nanobody, Orphan datasets	Outperformed AlphaFold2 (MSA) with better RMSD (e.g., 2.65 Å on IgFold-Ab) and TM-score of 0.43 on orphan dataset

Sr#	Study	Objective	ML method	Dataset	Key results
16	[19]	Enhance protein structure prediction with modern DL methods	CNNs, RNNs, LSTMs, Transformers	CASP , ProteinNet, PDB	Performance measured using TM and GDT scores; demonstrated improved structural predictions
17	[20]	Achieve highly accurate protein structure prediction using AlphaFold	AlphaFold, Neural Networks, Evoformer	PDB (up to Apr 2018), Uniclust30, BFD, MGnify, UniRef90	Reported 0.96 Å RMSD, with high pLDDT, GDT, and TM-scores; set new benchmarks in CASP14
18	[21]	Evaluate performance of single-sequence methods over MSA-based methods for PSP	AlphaFold2, ESMFold, OmegaFold, AminoBERT, RGN2	CASP15, RGN2, trRosettaX-Single publications	Single-sequence models scored TM > 0.7 for 24% of orphan proteins; designed sequences had higher median scores
19	[22]	Predict secondary structure using amide frequencies with data mining classification	Random Forest Classifier	Amide frequency dataset (alpha vs non-alpha structures)	Achieved high predictive accuracy with AUC of 0.963
20	[23]	Present TE-SS model using transformer encoder and Ankh PLM for PSSP	TE-SS model	PISCES, CASP12–14, CB433, TEST2016, TEST2018	Achieved Q3: 87.35%, Q8: 79.08%, Q9: 78.95%; SOV values over 76% across datasets
21	[24]	Propose MCP framework for accurate secondary structure prediction	MultiComponent Predictor, Fuzzy kNN, edit-SVM	RS126, CASP11, CASP12	MCP models achieved up to 87.29% accuracy; Fuzzy kNN with LZ scores: 76.38%; edit-SVM alone: 82.27%
22	[25]	Develop DeepPotential for high-accuracy PSP by integrating geometrical and spatial constraints	DeepPotential, Gradient Descent Folding	PDB, CASP13, CASP14, CAMEO (hard targets)	Achieved precision of 0.608 (CASP13), 0.687 (CAMEO); outperformed trRosetta; 6.7% higher TM-score in CASP14
23	[27]	Improve MULTICOM system performance in CASP14 using deep learning-based tools	DeepDist, DeepRank, DeepRank-con, DeepRank3-Cluster	CASP14	Reported performance using Sum Z-score, average TM-score, and average GDT-TS on CASP14 targets

Sr#	Study	Objective	ML method	Dataset	Key results
24	[29]	Use ML for improved sequence alignment in template-based PSP	Supervised ML, kNN for alignment scoring	SCOP, SCOP40	Achieved AUC up to 0.701 for most targets; TM-score improved to 0.499 for template alignments
25	[30]	Discover secondary protein structures using Local Euler Curvature and ML classifiers	rbLEC (LEC + RF Classifier), Unsupervised Clustering	CATH, LEC dataset	rbLEC model accuracy: 0.79 (vs. DSSP 0.74); F1-score: 0.86 (alpha), 0.93 (beta), 0.79 (coil) using STRIDE comparisons
26	[31]	Improve speed and accuracy of PSP for proteins lacking sequence/structure homologs	DeepMSA2, Deep Residual Networks	PDB, SCOPe 2.06, CASP9–12	TM-score 0.751 overall; 92.3% of proteins correctly folded; hard targets TM-score: 0.494 (> 40%)
27	[35]	Predict protein secondary structure using deep neural networks	CDNN, CNN, LSTM	CullPDB, CB513	Achieved Q3 accuracy of 77.5% on the CB513 dataset
28	[36]	Propose hybrid CNN-SVM model for improved PSSP performance	1D-CNN, SVM	CullPDB, CB513	Q8: 68.735%, Q3: 81.49%; CNN-SVM outperformed fine-tuned CNN by 0.024% in Q8 accuracy
29	[37]	Introduce ProteiNN: a transformer-based model for single-sequence PSP and DL trends review	ProteiNN (Transformer)	ProteinNet (CASP12 specifically)	Final RMSE: 0.448 (training), 0.4418 (validation), 0.4379 (test); end-to-end performance analyzed
30	[40]	Enhance PSSP using ensemble DL with NLP metrics and XAI techniques	DL model, BiLSTM, dense layers, XAI (LIME, Integrated Gradients)	PS4, CB513, CASP12, TS115	Accuracy: 94.41%, validation loss: 0.1585; supported by ROUGE-L structural validation
31	[41]	Develop MULTICOM2 for faster and more accurate PSP than its predecessor	DeepDist, DFOLD, trRosetta, DeepMSA, APOLLO, SBROD	CASP14	TM-score: 0.720 (first), 0.751 (best) on TBM domains; for FM domains: 0.514 and 0.540; 55–58% correct fold predictions
32	[42]	Propose a multi-scale CNN model for better feature extraction in PSSP	Convolutional Attention Neural Network	ASTRAL, CullPDB, CASP9–CASP12	90.12% accuracy on CASP10; 89.17% in 10-fold cross-validation; outperformed PSRSM and PSIPRED by 5.6% and 8.8% on CASP12

Sr#	Study	Objective	ML method	Dataset	Key results
33	[44]	Improve PSSP using models that capture local and long-range dependencies	BTCN, BLSTM, MSBTCN	CASP10–CASP14, CB513	Q3 and Q8 accuracy improved by 4.74% and 5.43% compared to prior models
34	[28]	Develop an accurate PSSP model using CNN combined with bidirectional GRU	Convolutional Bidirectional GRU (CBi-GRU)	CullPDB, CB513, CASP10, CASP11	Achieved 76.2% (CASP10) and 76.4% (CASP11); improved 2% over RaptorX-SS and 5.1% over DeepCNF
35	[45]	Develop constraint-guided conformational sampling for PSP	Constraint-guided neighbor generation, ML model (SPOT-1D)	Not specified	Improved average RMSD by ~ 1 Å over state-of-the-art search algorithms
36	[46]	Improve 8-state PSSP using a hybrid deep learning model	CRRNN (CNN + ResNet + Bi-GRU)	TR12148, CB513, CASP10–CASP12	Accuracy of 71.4% (CB513); eCRRNN ensemble improved to 74%
37	[47]	Present ThreaderAI for template-based prediction of tertiary structures	ResNet, Maximum Accuracy Algorithm, AdamW	SCOPe40, CASP13	TM-score 0.510 and GDT 0.437; outperformed HHpred, CNFpred, and CEThreader by 0.56, 0.13, and 0.11 in TM-score
38	[48]	Improve PSP using TCN-BiLSTM-MHA and knowledge distillation	TCN, BiLSTM, MHA	TS115, CB513, PDB (2018–2020), CASP13–CASP15	Q3 accuracy: +8.1%, SOV99: +25.9% (TS115); Q8 accuracy: +7.9% (CB513); MiAUC improved +1%
39	[49]	Enhance PSP through DeepMSA2 pipeline for multimeric structure prediction	DeepMSA2, modified AlphaFold2	Tara, MetaSource, JGIclust, CASP	In CASP15, outperformed AlphaFold2-Multimer; generated 47% correct complex structures
40	[50]	Improve PSSP accuracy using Hessian Free Optimization with FFNNs	Hessian Free Optimization, SVM	PISCES, CASP13	Achieved 80.4% Q3 and 0.77 SOV (PISCES); 78.14% Q3 and 0.755 SOV (CASP13)
41	[51]	Design CNN and LSTM models for PSSP from primary structures	CNN, LSTM	CulledPDB (PDB + PISCES)	Achieved Q3: 87.05% (CNN), 87.47% (LSTM); LSTM outperformed CNN slightly

Sr#	Study	Objective	ML method	Dataset	Key results
42	[43]	Accurately predict secondary structure using bio-inspired learning algorithms	Clonal Selection Algorithm (CSA), Multilayer Perceptron (MLP)	PDB	Accuracy improved from 89.41% to 96.61% by increasing CSA iterations and MLP epochs

Despite these achievements, several limitations persist. One of the most challenging areas remains the prediction of IDPs. These proteins do not adopt a fixed 3D conformation under physiological conditions and are involved in critical cellular functions such as signaling, regulation, and protein-protein interactions. Current DL models often underperform on IDPs due to their dynamic and context-dependent behavior. Integrating molecular dynamics simulations or physics-informed neural networks with ML techniques, as shown in the work by Heo and Feig [16], may help enhance predictive performance for such flexible protein regions.

Another challenge concerns dataset biases. Most DL models are trained on data from the PDB [34], which predominantly includes proteins that are easy to crystallize. This introduces bias toward globular and stable proteins, under representing membrane proteins, disordered regions, or protein complexes. Although the CASP challenges [32] and UniProt [38] provide more diverse datasets, the lack of annotations for edge cases continues to limit generalizability. Future datasets must ensure representation across diverse protein families, organisms, and structural types to avoid overfitting.

Moreover, the interpretability of deep learning models in structural biology is still a concern. Many DL architectures are "black boxes" making it difficult for researchers to trace how a specific prediction was made. This lack of transparency hinders trust and adoption in sensitive domains such as drug discovery and clinical genomics. To address this, the integration of XAI methods such as SHAP, LIME, and Integrated Gradients is gaining attention. For example, Vignesh et al. [40] demonstrated how integrating LIME with BiLSTM models could provide insights into which sequence motifs contributed most to the structural output. This trend is likely to grow, with future models offering interpretable predictions, especially in regulatory settings.

Additionally, hybrid models that combine classical ML algorithms with deep learning frameworks have shown strong potential in capturing different protein features. For instance, SERT-StructNet [10] integrates CNNs, BiLSTMs, SENet modules, and transformers to improve secondary structure prediction, outperforming traditional models like DeepCNF and RaptorX-SS. Similarly, ensemble methods using SVMs, RF, and kNN have been employed to refine classification outputs and improve accuracy [12]. These hybrid strategies allow models to leverage both statistical and learned representations, enhancing both performance and robustness.

Another emerging trend is the combination of ML with biophysical constraints. For example, DeepPotential [25] and ThreaderAI [47] incorporate residue-residue distance potentials and template threading data to guide model learning. This not only improves accuracy but also ensures that predictions are biologically plausible, which is essential for downstream applications such as rational drug design, protein engineering, and synthetic biology.

In summary, while ML has significantly advanced protein structure prediction, the field must

still address several open challenges. These include improving the modeling of disordered and complex proteins, reducing training biases, increasing model interpretability, and fostering multi-disciplinary integration between AI and structural biology. The future of PSP lies in hybrid architectures, dataset diversification, and transparent AI frameworks. Ultimately, these innovations will bridge the gap between computational predictions and real-world biological functions, paving the way for applications in precision medicine, personalized therapeutics, and beyond.

Conclusion

Machine learning, particularly deep learning, has revolutionized protein structure prediction, with groundbreaking models like AlphaFold and HelixFold-Single setting new standards for accuracy and efficiency. These advancements have far-reaching implications for drug discovery, synthetic biology, and disease modeling, enabling researchers to explore the molecular underpinnings of complex diseases and design targeted therapeutic interventions.

The integration of evolutionary data, physical constraints, and advanced computational techniques has significantly enhanced the predictive capabilities of ML models. However, challenges such as the accurate modeling of IDPs, dataset biases, and the interpretability of ML models remain. Addressing these issues will require continued innovation in quantum mechanics, physics-based simulations, and XAI techniques.

As the field progresses, the refinement of ML methodologies and the development of more diverse and representative datasets will be critical. These efforts will not only improve prediction accuracy but also unlock new opportunities in precision medicine, rational drug design, and biotechnological innovations. The future of protein structure prediction lies in the seamless integration of computational and experimental approaches, paving the way for a deeper understanding of biological systems and their applications in solving real-world problems.

Glossary of terms

NMR	Nuclear Magnetic Resonance
ML	Machine Learning
PSP	Protein Structure Prediction
kNN	k-Nearest Neighbor
SCOP	Structural Classification of Proteins
RMSD	Root-Mean-Square Deviation
pLDDT	Predicted Local-Distance Difference Test
GDT	Global Distance Test
TM-Score	Template Modeling Score
NN	Neural Network
PDB	Protein Data Bank
AUC	Area Under Curve
DL	Deep Learning
MSAs	Multiple Sequence Alignments
FFNNs	Feedforward Neural Networks
RNNs	Recurrent Neural Networks
CNNs	Convolutional Neural Networks
D-CNNs	Deep Convolutional Neural Networks

BRNNs	Bidirectional RNNs
PSSP	Protein Secondary Structure Prediction
SVM	Support Vector Machine
PLM	Protein Language Model
CDNN	Combining Deep Neural Networks
LSTM	Long Short-Term Memory Networks
CASP	Critical Assessment of Protein Structure Prediction
TM-scores	Template Modeling Scores
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
SOV	Share of Voice
MAE	Mean Absolute Distance Error
TCN	Temporal Convolutional Network
BiLSTM	Bidirectional Long Short-Term Memory
MHA	Multi-Head Attention
PSS	Protein Secondary Structure
DT	Decision Tree
NB	Naïve Bayes
RF	Random Forest
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
NLP	Natural Language Processing
RGN2	Recurrent Geometric Network
GNNs	Graph Neural Networks
DLMs	Deep Learning Models
BLS	Broad Learning System
PISCES	Protein Sequence Culling Server

References

1. Afify H. M., M. B. Abdelhalim, M. S. Mabrouk, A. Y. Sayed (2021). Protein Secondary Structure Prediction (PSSP) Using Different Machine Algorithms, Egyptian Journal of Medical Human Genetics, 22(1), 54.
2. Anfinsen C. B. (1973). Principles that Govern the Folding of Protein Chains, Science, 181(4096), 223-230.
3. Asgari E., N. Poerner, A. C. McHardy, M. R. K. Mofrad (2019). DeepPrime2Sec: Deep Learning for Protein Secondary Structure Prediction from the Primary Sequences, bioRxiv, 705426.
4. Baek M., F. DiMaio, I. Anishchenko, J. Dauparas, et al. (2021). Accurate Prediction of Protein Structures and Interactions Using a Three-track Neural Network, Science, 373(6557), 871-876.
5. Bryant P. (2023). Deep Learning for Protein Complex Structure Prediction, Current Opinion in Structural Biology, 79, 102529.
6. Burley S. K., H. M. Berman (2021). Open-access Data: A Cornerstone for Artificial Intelligence Approaches to Protein Structure Prediction, Structure, 29(6), 515-520.
7. Chandni K., M. Pandya, S. Jardosh (2018). Deep Learning Approaches for Protein Structure Prediction, International Journal of Engineering & Technology, 7(4.5), 168-170.
8. Charalampous K., M. Agathocleous, C. Christodoulou, V. Promponas (2022). Solving the Protein Secondary Structure Prediction Problem with the Hessian Free Optimization Algorithm, IEEE Access, 10, 27759-27770.

9. Chowdhury R., N. Bouatta, S. Biswas, C. Floristean, et al. (2022). Single-sequence Protein Structure Prediction Using a Language Model and Deep Learning, Nature Biotechnology, 40(11), 1617-1623.
10. Dong B., Z. Liu, D. Xu, C. Hou, et al. (2024). Sert-Structnet: Protein Secondary Structure Prediction Method Based on Multi-Factor Hybrid Deep Model, Computational and Structural Biotechnology Journal, 23, 1364-1375.
11. Elnaggar A., M. Heinzinger, C. Dallago, G. Rehawi, et al. (2021). ProtTrans: Toward Understanding the Language of Life Through Self-supervised Learning, IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112-7127.
12. Ema R. R., M. A. Khatun, M. N. Adnan, S. S. Kabir, et al. (2022). Protein Secondary Structure Prediction Based on CNN and Machine Learning Algorithms, International Journal of Advanced Computer Science and Applications, 13(11), 102844.
13. Fang X., F. Wang, L. Liu, J. He, et al. (2023). A Method for Multiple-sequence-alignment-free Protein Structure Prediction Using a Protein Language Model, Nature Machine Intelligence, 5(10), 1087-1096.
14. Girgis M. R., R. M. Gamal, E. Elgeldawi (2022). Ensemble Machine Learning to Enhance Q8 Protein Secondary Structure Prediction, Computers, Materials and Continua, 73(2), 3951-3967.
15. Guo Y., W. Li, B. Wang, H. Liu, et al. (2019). DeepACLSTM: Deep Asymmetric Convolutional Long Short-term Memory Neural Models for Protein Secondary Structure Prediction, BMC Bioinformatics, 20(1), 341.
16. Heo L., M. Feig (2020). High-accuracy Protein Structures by Combining Machine-learning with Physics-based Refinement, Proteins, 88(5), 637-642.
17. Jayasimha A., R. Mudambi, P. Pavan, B. M. Lokaksha, et al. (2021). An Effective Feature Extraction with Deep Neural Network Architecture for Protein-secondary-structure Prediction, Network Modeling Analysis in Health Informatics and Bioinformatics, 10(1), 58.
18. Jing X., F. Wu, X. Luo, J. Xu (2023). RaptorX-single: Single-sequence Protein Structure Prediction by Integrating Protein Language Models, bioRxiv, 2023-04.
19. Jisna V. A., P. B. Jayaraj (2021). Protein Structure Prediction: Conventional and Deep Learning Perspectives, The Protein Journal, 40(4), 522-544.
20. Jumper J., R. Evans, A. Pritzel, T. Green, et al. (2021). Highly Accurate Protein Structure Prediction with Alphafold, Nature, 596(7873), 583-589.
21. Kandathil S. M., A. M. Lau, D. T. Jones (2023). Machine Learning Methods for Predicting Protein Structure from Single Sequences, Current Opinion in Structural Biology, 81, 102627.
22. Kathuria C., D. Mehrotra, N. K. Misra (2018). Predicting the Protein Structure Using Random Forest Approach, Procedia Computer Science, 132, 1654-1662.
23. Kazm A. A., A. Ali, H. Hashim (2024). Transformer Encoder with Protein Language Model for Protein Secondary Structure Prediction, Engineering, Technology & Applied Science Research, 14(2), 13124-13132.
24. Khalatbari L., M. R. Kangavari, S. Hosseini, H. Yin, et al. (2019). MCP: A Multi-Component Learning Machine to Predict Protein Secondary Structure, Computers in Biology and Medicine, 110, 144-155.
25. Li Y., C. Zhang, D.-J. Yu, Y. Zhang (2022). Deep Learning Geometrical Potential for High-accuracy *ab initio* Protein Structure Prediction, iScience, 25(6), 104425.
26. Lin Z., H. Akin, R. Rao, B. Hie, et al. (2022). Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction, bioRxiv, 500902.
27. Liu J., T. Wu, Z. Guo, J. Hou, et al. (2022). Improving Protein Tertiary Structure Prediction by Deep Learning and Distance Prediction in CASP14, Proteins: Structure, Function, and Bioinformatics, 90(1), 58-72.

28. Lu Y. (2024). Protein Secondary Structure Prediction Using Convolutional Bidirectional GRU, Journal of Mathematics Research, 16(4), 11.
29. Makigaki S., T. Ishida (2019). Sequence Alignment Using Machine Learning for Accurate Template-based Protein Structure Prediction, Bioinformatics, 36(1), 104-111.
30. Moreira, R. A., R. Braddell, F. A. Santos, T. Fülöp, et al. (2023). Discovering Secondary Protein Structures via Local Euler Curvature, bioRxiv, 2023-11.
31. Pearce R., Y. Li, G. S. Omenn, Y. Zhang (2022). Fast and Accurate *ab initio* Protein Structure Prediction Using Deep Learning Potentials, PLoS Computational Biology, 18(9), e1010539.
32. Protein Structure Prediction Center (2026). Critical Assessment of Techniques for Protein Structure Prediction (CASP), https://predictioncenter.org/ (access date 08 June 2026).
33. Rao R., J. Liu, R. Verkuil, J. Meier, et al. (2021). MSA Transformer, Proceedings of the International Conference on Machine Learning, 8844-8856.
34. RCSB Protein Data Bank (2026). https://www.rcsb.org (access date 08 June 2026).
35. Suh D., J. W. Lee, S. Choi, Y. Lee (2021). Recent Applications of Deep Learning Methods on Evolution- and Contact-based Protein Structure Prediction, International Journal of Molecular Sciences, 22(11), 6032.
36. Sutanto V. M., Z. I. Sukma, A. Afiahayati (2021). Predicting Secondary Structure of Protein Using Hybrid of Convolutional Neural Network and Support Vector Machine, International Journal of Intelligent Engineering and Systems, 14(1), 232.
37. Szelogowski D. (2023). Deep Learning for Protein Structure Prediction: Advancements in Structural Bioinformatics, bioRxiv, 2023-04.
38. The UniProt Consortium (2023). UniProt: The Universal Protein Knowledgebase in 2023, Nucleic Acids Research, 51(D1), D523-D531.
39. The UniProt Consortium (2026). UniProt: the Universal Protein Knowledge Base, https://www.uniprot.org (access date 08 June 2026).
40. Vignesh U., R. Parvathi, K. G. Ram (2024). Ensemble Deep Learning Model for Protein Secondary Structure Prediction Using NLP Metrics and Explainable AI, Results in Engineering, 24, 103435.
41. Wu T., J. Liu, Z. Guo, J. Hou, et al. (2021). MULTICOM2 Open-Source Protein Structure Prediction System Powered by Deep Learning and Distance Prediction, Scientific Reports, 11(1), 13155.
42. Xu Y., J. Cheng (2021). Secondary Structure Prediction of Protein Based on Multi-Scale Convolutional Attention Neural Networks, Mathematical Biosciences and Engineering, 18(4), 3404-3422.
43. Yavuz B. Ç., N. Yurtay, O. Ozkan (2018). Prediction of Protein Secondary Structure with Clonal Selection Algorithm and Multilayer Perceptron, IEEE Access, 6, 45256-45261.
44. Yuan L., X. Hu, Y. Ma, Y. Liu (2022). DLBLS_SS: Protein Secondary Structure Prediction Using Deep Learning and Broad Learning System, RSC Advances, 12(52), 33479-33487.
45. Zaman R., M. A. H. Newton, F. Mataeimoghadam, A. Sattar (2022). Constraint Guided Neighbor Generation for Protein Structure Prediction, IEEE Access, 10, 54991-55001.
46. Zhang B., J. Li, Q. Lü (2018). Prediction of 8-State Protein Secondary Structures by a Novel Deep Learning Architecture, BMC Bioinformatics, 19(1), 293.
47. Zhang H., Y. Shen (2020). Template-Based Prediction of Protein Structure with Deep Learning, BMC Genomics, 21(Suppl 11), 878.
48. Zhao L., J. Li, W. Zhan, X. Jiang, et al. (2024). Prediction of Protein Secondary Structure by the Improved TCN-BiLSTM-MHA Model with Knowledge Distillation, Scientific Reports, 14(1), 16488.

49. Zheng W., Q. Wuyun, Y. Li, C. Zhang, et al. (2024). Improving Deep Learning Protein Monomer and Complex Structure Prediction Using DeepMSA2 with Huge Metagenomics Data, Nature Methods, 21(2), 279-289.
50. Zhou S., H. Zou, C. Liu, M. Zang, et al. (2020). Combining Deep Neural Networks for Protein Secondary Structure Prediction, IEEE Access, 8, 84362-84370.
51. Zubair M., M. K. Hanif, E. Alabdulkreem, Y. Ghadi, et al. (2022). A Deep Learning Approach for Prediction of Protein Secondary Structure, Computers, Materials and Continua, 72(2), 3705-3718.

Assist. Prof. Hassan Tariq, Ph.D.
E-mail: hassan@uaf.edu.pk

Hassan Tariq got his Ph.D. in Computer Sciences from the Victoria University of Wellington, New Zealand. Currently, he is an Assistant Professor in the Department of Computer Science, University of Agriculture, Faisalabad, Pakistan. He has more than 15 years of teaching and research experience. He is the head of various departmental committees. He is also a convener of stakeholders and Industrial linkages. Dr. Tariq served as an examiner of various M.Sc. and Ph.D. thesis.

Areej Fatima, B.Sc.
E-mail: areejfatima7644@gmail.com

Areej Fatima holds a B.Sc. degree in Bioinformatics. She has expertise in structural bioinformatics, with a focus on applying machine learning and deep learning techniques to address challenges in the biological sciences. Her work involves leveraging computational tools to analyze protein structures, predict molecular interactions, and contribute to advancements in drug discovery and disease modeling.

Muhammad Sohaib, B.Sc.
E-mail: sohaibafzal102@gmail.com

Muhammad Sohaib holds a B.Sc. degree in Bioinformatics. His research is centered on the application of machine learning techniques to solve complex problems in the biological sciences. With a strong focus on computational biology, he explores innovative approaches to protein structure prediction, molecular modeling, and the analysis of biological datasets, contributing to advancements in precision medicine and therapeutic development.