Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 32 No. 10s (2025)

Role of Deep Learning in Protein Structure Prediction: Current Progress and Open Challenges

Bhushan Wakode[1], Dr S P Deshpande[2]

[1]Research Scholar, PGDCST, DCPE, Amravati, India

[2]Professor, PGDCST, DCPE, Amravati, India

E-mai:bhushan.wakode@gmail.com

Article History:

Received: 12-01-2025

Revised: 15-02-2025

Accepted: 01-03-2025

Abstract:

Introduction: Protein structure prediction is a fundamental challenge in computational biology, with significant implications for drug discovery, disease modelling, and biotechnology. Traditional experimental methods such as X-ray crystallography, nuclear magnetic resonance spectroscopy, and cryo-electron microscopy provide highly accurate structural insights but are expensive, time-consuming, and often unsuitable for complex or membrane-bound proteins. To address these limitations, computational approaches have emerged, categorized into physics-based simulations, evolutionary-based modeling, and deep learning techniques. Recent advancements in machine learning and deep learning have revolutionized protein structure prediction. Models like AlphaFold2 and RoseTTAFold have achieved near-experimental accuracy by leveraging large-scale protein datasets and advanced neural network architectures. However, challenges remain in multi-protein interaction modeling, side-chain conformation prediction, dataset biases, and computational efficiency. Additionally, the black-box nature of deep learning models limits their interpretability, necessitating efforts to enhance transparency and explainability in AI-driven protein modeling.

Objectives: The main objective of this study is to comprehensive review the protein structure prediction methodologies, highlighting the evolution from traditional computational models to cutting-edge deep learning frameworks.

Methods: It discusses key challenges, including dataset limitations, model scalability, and integration with experimental techniques, and explores future research directions such as self-supervised learning, quantum computing for protein folding simulations, and energy-efficient deep learning architectures.

Conclusions: By addressing these challenges, computational protein modeling can further advance biomedical research, enabling more accurate disease modeling, rational drug design, and the development of synthetic biomolecules with tailored functions.

Keywords: Amino Acid Sequence, Protein structure prediction, multi-protein interaction, Protein sequence, drug design.

1. Introduction

Proteins are fundamental to nearly all biological functions, serving as enzymes, structural components, and signaling molecules within cells. Unlike DNA, which primarily stores genetic information, proteins actively participate in cellular processes through dynamic interactions [1]. The

3D structure of a protein determines its function, influencing its ability to bind to other molecules, catalyze biochemical reactions, and regulate various physiological mechanisms [2]. Understanding protein structure is crucial in fields such as drug discovery, disease modeling, and biotechnology, where identifying functional proteins can lead to novel therapeutic interventions [3]. Despite their biological significance, determining protein structures remains a challenging task. Traditional experimental techniques such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM) provide highly accurate structural insights but are expensive, time-consuming, and not always feasible for all proteins [4]. Many proteins, especially membrane-bound or intrinsically disordered proteins, are difficult to crystallize or analyze using these methods [5]. The increasing demand for faster and more cost-effective alternatives has driven the development of computational approaches to predict protein structures with high accuracy.

Computational methods for protein structure prediction can be broadly categorized into physics-based and evolutionary-based approaches. Physics-based models simulate protein folding based on molecular interactions, energy functions, and force fields, attempting to replicate the folding process from first principles [6]-[7]. These methods, while theoretically sound, require extensive computational power and often struggle with complex proteins. In contrast, evolutionary-based approaches leverage sequence homology and statistical patterns from known protein databases to infer structural information [8]. These techniques are highly effective when similar protein sequences are available but may fail for novel proteins lacking homologous templates [9].

The emergence of machine learning and deep learning has significantly transformed protein structure prediction. By training neural networks on vast protein datasets, these models can capture intricate patterns in amino acid sequences and structural properties. The Critical Assessment of Structure Prediction (CASP) competition has been instrumental in benchmarking advancements in this field [10]-[11]. The breakthrough success of DeepMind's AlphaFold2, which achieved near-experimental accuracy in CASP14 with a Global Distance Test (GDT) score of approximately 90%, highlights the potential of deep learning (DL)-driven methods [12]. However, challenges remain, particularly in predicting multi-protein interactions, modeling flexible regions, and incorporating environmental factors such as post-translational modifications. A typical structure of protein is shown in Figure 1. Numerous fields approach the challenge of structuring proteins prediction. Because it is a prerequisite for successfully battling fresh proteins models, researchers were captivated by the framework prediction problem [13]. Creating protein patterns that coil into proteins that carry out the desired functions is the ultimate goal of completely novel protein structure. Within the molecular level, entirely from new protein structure can be thought of like a design-related problem.

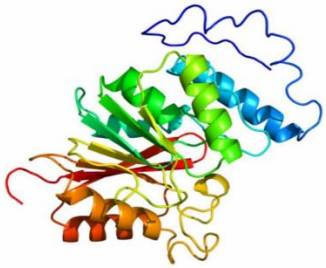

Figure 1: Typical Structure of Protein [1]

The theoretically maximum of the 3-class secondary structure of proteins predictions is being approached by cutting-edge techniques, which have recently achieved accuracy scores of nearly 89%. As a result, several academics begin focusing on more difficult tasks, such as developing multi-task categories for predictions or predicting the 8-class secondary framework. Employing sequential and historical data, Malhotra S. et al. trained one artificial neural network for predicting several protein attributes, such as solvent availability and protein secondary structure [14]. Three different types of structural features are predicted by a deep neural network created by Heffernan et al. [15]: solvents exposed surface space, torsional viewpoints, C-atom related angles and the dihedral viewpoints, and secondary structure. The quantity of subsequent amino acid sequence predictions articles is displayed in Figure 2.

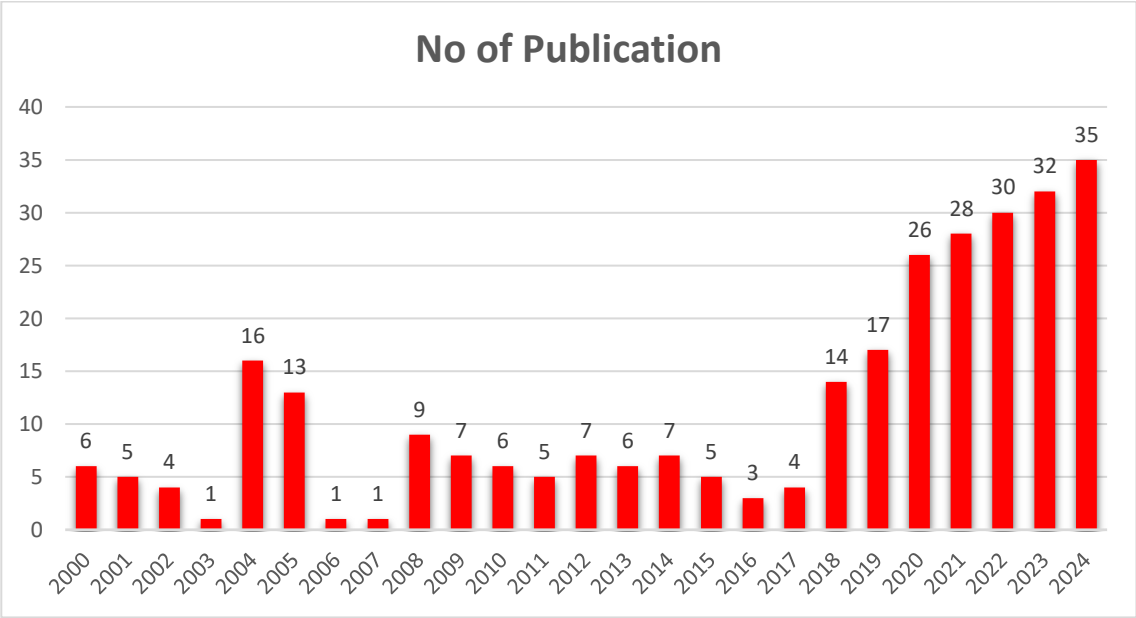

Figure 2: Year wise no of Publications of secondary protein sequence prediction

This review provides a comprehensive analysis of protein structure prediction methods, tracing the evolution from traditional physics-based simulations to cutting-edge deep learning models. It explores the strengths and limitations of existing techniques, the role of artificial intelligence in overcoming prediction challenges, and potential future directions in protein modeling. By addressing the critical gaps in current methodologies, this study aims to contribute to the ongoing advancements in structural bioinformatics and computational biology.

2. Search Strategy

To identify relevant research on secondary protein sequence prediction, a systematic search was conducted using multiple databases, including IEEE Xplore, PubMed, Scopus. The search spanned papers published between 2015 and 2025, focusing on keywords such as "Protein sequence," "Secondary protein sequence," "Artificial Intelligence," "Deep Learning," and "Machine Learning." Inclusion criteria prioritized studies that presented novel AI-based methodologies, used well-established datasets, or introduced significant advancements in abnormal event detection. Exclusion criteria filtered out works unrelated to surveillance systems, non-AI-based methods, or studies

lacking experimental validation. A meticulous review of abstracts and full texts ensured the selection of high-quality studies, which were further categorized based on their techniques.

Table 1: Keyword based criteria

Basic Keywords	**"**Protein sequence**"**			
"Direct Keywords"	"Deep Learning"	"Machine Learning"	"Artificial Intelligence"	
"Indirect Keywords"	"Secondary Protein Sequence"	"Amino Acid Sequence"	"Peptide prediction"	"Protein Secondary Sequence prediction"

Inclusion and Exclusion Criteria

A number of analysis processes were created to select pertinent academic publications for the research evaluation and eliminate criteria for investigation exclusions.

The screening process uses three terms for inclusion standards.

Inclusion: Papers published between 2015–2025 focusing on AI-based methods.

Exclusion: Studies not directly related to Secondary Protein Sequence prediction.

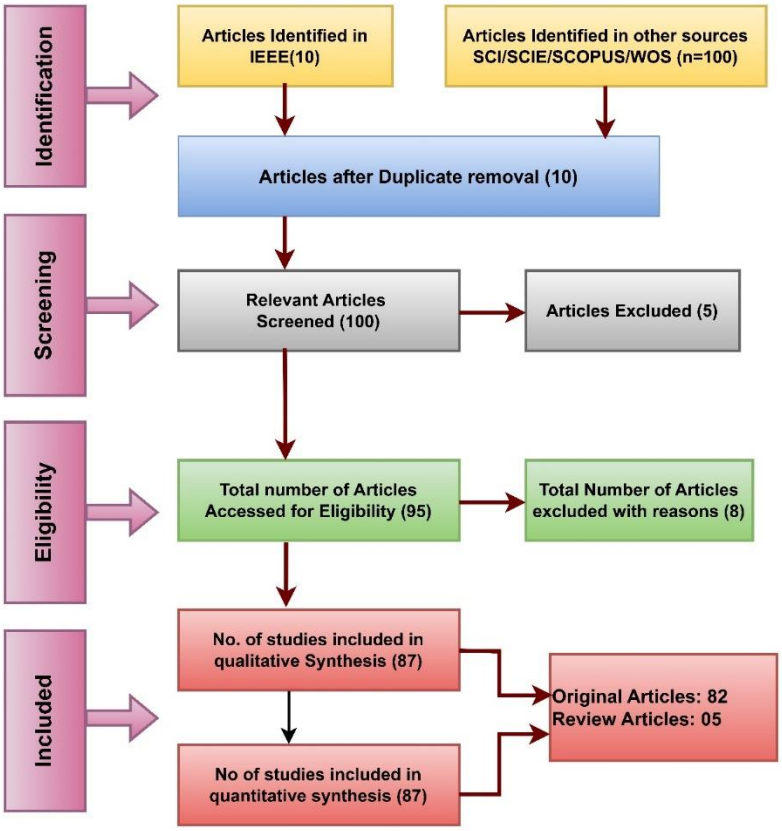

Figure 3: PRISMA Approach

3. Background of Protein Structure

There is a computational strategy which can be broadly classified into two groups: (a) those that employ physical theories and (b) those depending on evolutionary concepts. The evolution of predicting protein structure across the two groups using different statistical ML and DL techniques. Employing molecular interactions based on a force area's prospective energy at a specific moment or fragmentation assembling employing the power function to create a power-stable three-dimensional framework, the physics-based method mimics the folding operation of amino acid chains. However, fragments assembly gets reasonable precision if it includes knowledge about protein comparable, whereas molecular structure works best for tiny proteins [16].

An evolutionary theory that is predicated on the idea that every organism shared a predecessor and subsequently developed as a result of environmental adaption. The protein structure gets modified in this modification to enable the best possible function. The necessary amino acids remain unchanged whenever the structure changes. DNA is the process wherein only the vital and enabling amino acids alter to conform to the physicochemical conditions whereby the protein needs exist. Nevertheless, this method necessitates comparable sequence data, making it challenging to ascertain the structure of an entirely novel protein.

It is also challenging to examine how variations affect function using this method.

Physic-based Methods: Creating energy functions that direct the behavior of proteins in the conformational domain as it moves from unfolding to the folding phase is a common practice in physics-based approaches to protein folding. The first principle atomistic force field [17]-[18] was used in a number of methods over the past several decades to develop performance measures [19]-[20], and protein modeling [21]-[22] to establish a coarse-grained method. In this situation, performance measures that take into consideration multibody factors which are difficult to predict theoretically can be designed with the use of neural network models. The model is trained on amino acid strands having any knowledge of their architecture in order to predict the interactions across acids in an unsupervised manner. Direct Coupling Investigation, a methodology that uses Markov Random Fields toward the MSA underneath a protein strand, is the main method for examining how evolution determines limits among collections of comparable protein sequences. A number of research have suggested replacing shallow Markov Random Fields (MRF) model. Riesselman et al. demonstrated that protein functionality may be determined from inconsistent sequences by training an autoregressive model over MSA while ignoring sequence [23]. Instead of using a complete framework for obtaining interaction proteins, Riesselman et al. [23] only employed a set of associated segments, relative to study by Rao et al. [24], that employed numerous MSA. To create 3D models, a different approach employs the LSTM framework with information within the shape of protein sequences, PSSM, and torsional of distortion [16]. Using the findings of Alquraishi's research [25] on the geometrical phases, a framework was constructed with three stages: computational, geometric, and assessment. Developing performance parameters that direct the movement of proteins in the conformational environment from the unfolding to the folding phase is a common practice in physics-based methods towards protein folding. In recent decades, a number of methods have been used to develop energy functions [19]-[20], and through the use of the fundamental fragmented

power fields [26]-[27]. These methods were thereafter modified by the use of protein modeling [21]-[22], and a course in texture technique.

In this situation, performance measures that take into consideration multibody factors that are difficult to predict analytical can be designed with the use of computational neural networks. The framework is trained on amino acid strands lacking any knowledge of protein shape in order to predict the interaction across acids in a detached way. Direct Connectivity Investigation, a methodology that uses Markov Random Fields toward the MSA underneath an amino acid strand, is the main method for examining the developmental limits among collections of comparable proteins sequences. A number of research have suggested replacing shallow Markov Random Fields with DNN model. Riesselman et al. demonstrated the possibility that protein function may be determined from inconsistent sequences by training a self-regressive approach towards MSA while ignoring alignment [23].

Instead of using a complete structure for obtaining interaction amino acids, Riesselman et al. [23] only employed a collection of associated strands, compared to study by Rao et al. [24], that utilized numerous MSA. To create 3D structures, a different approach employs the LSTM framework and takes as inputs amino acid chains [16].

Table 2: Compare the existing methods for protein sequence

Ref	**Methods**	**Contributions**
[28]-[29]-[30]	RGN-BERT, ProteinBERT, ProteinMPNN	Amino Acid sequence
[31]-[32]	1D-CNN, SVM, RF, CNN	Predicting Interaction
[33]-[34]	LOMETS3, PSA	Forecasting of Profiles
[35]	CASP14	Proximity Forecasting
[36]-[37]-[38]-[49]-[50]	Deep CNN, LSTM, StructNet, ResNet, Transformer	secondary structure
[39]-[40]-[41]	CASP13 and CASP14, ResNet, DLPacker	Sidechain Prediction
[42]	Ensemble	Tertiary Framework
[43]-[44]-[45]	DEEPCON, lDDT	Analysis and Evaluation of Prediction Models
[46]-[47]	GAN, RNN	Correctness Forecast
[48] [49]	C-CNN	Distance Forecast Forecasting

Evolutionary-based Methods: The application of a DNN approach has led to a significant advancement in protein structure detection [51]-[52]. The first controlled investigations utilized co-evolutionary traits [51]-[53]. Furthermore, MSA served as the data source for a trained protein structural identification approach. In [54]-[55] suggested into an architecture which analyzed MSA data instantly utilizing 2D convolutional layers [54] and a GRU [55]. In the most recent sophisticated polypeptide prediction effort, AlphaFold2, attentiveness was used to assess MSA-MSA using a supervised full structure of protein construction [56].

Protein interaction prediction represents a crucial component in every one of presently available cutting-edge protein structural predictions techniques [57]-[58] it plays a significant role in computerized protein development [59]-[60]. Numerous study teams have used an autonomous technique for molecular language modeling [61]-[62]. Transformers design was employed in a number of these research [61]-[62]-[63]. Utilizing data in the shape of statistics such as correlation [44]-[64], estimated convolutional features [65], patterns, or historical characteristics [66], DL techniques have demonstrated effectiveness when identifying interaction proteins [44]-[58]-[64]. Co-evolutionary component inclusion is crucial for modelling success, according to another research [64]. This method has proved applied in other fields, such as predicting protein structures [67]-[68] particularly the development of large-scale modeling of languages for processing natural words [69]. The sequence of terms in phrases is said to be comparable to the strands of amino acids that make up proteins, where a word may have an association between a word located adjacent to it as well as a phrase which is somewhat distant from its original location. Earlier investigations use a ML algorithm in predicting protein contact. Bepler and Berger created sequence integration by combining structure.

Researchers became initially able to pair associations among residues by fine-tuning a pre-trained employing a LSTM framework for protein sequences [70]. A.Rives demonstrated that characteristics of biology could be captured by the LSTM framework for language [71].

Through executing a controlled linear projection of protein structure, the initial attempt to use the linguistic model Transformers to analyze the protein structure demonstrated that data regarding the interaction among acids could be retrieved from the learnt model [71]. The self-attention process in Transformers was thoroughly examined in a further investigation. It discovered that the different layers in the framework assist in the investigation of different properties and determined its relevance to pertinent aspects of biology [72].

Rao et al. [73] contrast Transformers, which were developed upon a vast text database, with the Potts approach, which was learned with single MSA. The outcome demonstrates that transformers express interactions immediately via their paired element (self-attention), similar to how the Potts paradigm does by way of its paired components (weights). This work also demonstrates the connection among linguistic model confusion, MSA dimension, and model efficiency.

A number of additional research have investigated solutions to masking language modeling, including the use of a set of strands for supervising [74]-[75] conditioned synthesis [76], and a contrasting loss function approach [50]. Position-specific scored matrix pattern prediction using

unsupervised modeling of languages was extended by Sturmfels et al. [74]. Depreciated optimization was employed by Sercu et al. to concurrently forecast profiles and bilateral connectivity [75].

In recent years, MSA fits have been carried out using a DL technique. Heinzinger et al. demonstrated that there could be a correlation between protein structure and the variables examined using the VAE modelling [62]. Smith et al. optimized the overall architecture employing Rosetta and predicted the paired proximity employing the DRN by utilizing the Potts algorithm's properties with pseudolikelihood maximizing [77]-[78].

Transition and Comparative Analysis: While physics-based methods focus on fundamental molecular interactions, they often require extensive computational resources. In contrast, evolutionary-based methods rely on available protein sequence data but may struggle with novel proteins lacking homologs. Hybrid approaches combining both techniques—such as deep learning-assisted molecular simulations—are emerging as promising solutions to bridge the gap between these two paradigms.

4. Protein Structural Predictions utilizing ML and DL Approach

Protein frameworks, in contrast, are frequently difficult to establish via experimentation. However, using gene data has recently led to significant advancements. It has been feasible to determine which amino acid outcomes have previously been in communication through examining relationship in relatives, which aid in structuring proteins forecasting services. We demonstrate how artificial neural networks has been developed for predicting distances across relationships of contributions, that additionally offer more fundamental data than association forecasts Protein framework forecasting represents a technique for prediction a substance's 3-D shape determined by which is amino acid linkage. This becomes a crucial test because the structural makeup of a protein determines its functionality to an extensive dimension.

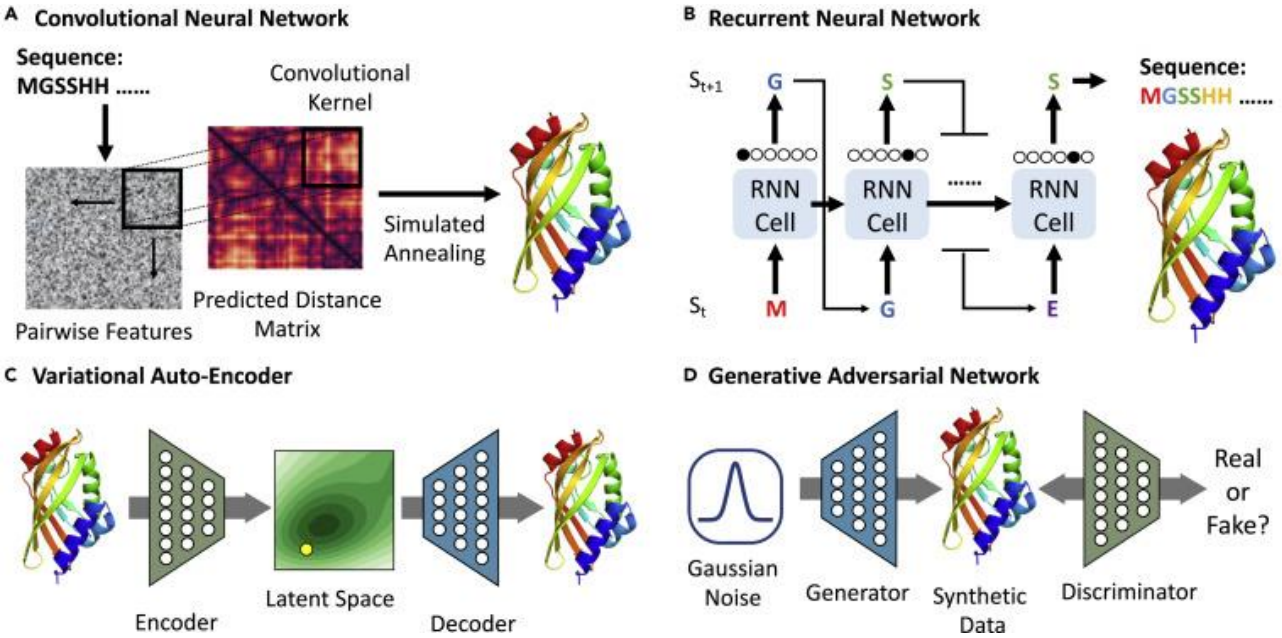

Figure 3: DL-based protein structure prediction [79]

To develop commercial strains capable of excessive production bioproducts, physiological engineering requires a thorough understanding of cellular metabolism, including metabolic processes and catalysts [80]. A significant obstacle to their biosynthesis is revealed by the fact that the metabolic pathways and enzymes involved in many molecules of significance remain unknown. By designing new biochemical routes and enzymes, this issue may become partially resolved. With the growth of bio-big information, data-driven approaches that make use of AI tools enable very complex protein and network models. The present study on AI-assisted molecular production and design has likely focused on controlled advancement, which also makes use of AI to rapidly generate mutation library.

Although protein framework forecasting represents a complicated issue, it is further frequently divided into and addressed at multiple phases: 1-D prediction of structural features together the crucial amino acid pattern; 2-D forecasting of geographical interaction between amino acids; 3-D prediction of a protein's secondary development; and 4-D forecasting of a multi-protein complex's quaternary configuration. AI techniques have been used frequently in information science, mathematical biology, and structures science. The development of ML approaches to protein framework forecasting appears to have significant effects in both bioinformatics and comprehensive biology [81].

Because of the importance of the observed challenge, its established physics, and its statistical foundation, peptide structure prediction from sequencing has in fact been the subject of numerous studies for years [82]. Although progress has fluctuated over decades, the "neutralization" of structural forecasting algorithms has advanced significantly in the earlier decade, including neural networks replacing algorithms that previously relied on energy models and sampled processes. With this research, a DNN mechanism for detecting protein-ligand interactions effectively when using the medication important had been developed [82]. The DNN determines which medication creates the most effective virus-fighting response and recognizes protein-ligand interactions for a certain medication.

Protein-ligand associations for a single medicine are effectively shown by the DNN using a restricted genome assembly of Indian individuals submitted to the GISAID dataset [83]. In subsequent decades, the study may be expanded to include a combination of DL techniques to thoroughly evaluate the peptide sequence resulting from protein-ligand relationships.

Given their popularity, clustering-dependent algorithms fail to identify good/near-native imitations on historical data when structural protein forecasting approaches appear to drastically under-sample near-native imitations [84]. However, it takes much longer to gather reservoirs from the ground employing these methods. An inventive deceptive techniques selecting method based on non-matrix decomposition is proposed in this work. It proves that the suggested approach outperforms the energy landscape-based approach. The time complexity issue and the challenges of locating suitable deception in an extremely sparse dataset are both addressed by this same proposed approach, which successfully identifies near-native deception across simple and difficult protein receptors.

The understanding of amino acid design has already advanced significantly as a result of the PDB's enormous collection of 3D structure data, which culminated in the most recent developments in

structural protein predictions supported by AI technologies [85]. Performance improves and the impact of AI advancements continues to expand as the amount of publicly available amino acid structure data rises. Using comprehensive or combination methods to effectively manage and restore information generated by structural investigations of ever-larger frameworks—from molecular tools to organs to entire cells—seems to have become the most significant task facing biological structures.

5. Challenges in protein Prediction

Despite remarkable advancements in computational protein structure prediction, several challenges persist, limiting the accuracy and applicability of existing models. These challenges stem from data limitations, model interpretability, computational costs, and the complexity of protein interactions. Addressing these issues is crucial for improving prediction reliability, especially in drug design, enzymatic engineering, and disease-related research.

Limitations of Datasets and Sequence Diversity: Deep learning models rely heavily on large, high-quality datasets to achieve accurate protein structure predictions. However, biases in protein databases can influence model performance. Publicly available datasets, such as Protein Data Bank (PDB) and UniProt, are often biased toward well-studied proteins, particularly those that are easier to crystallize and analyze experimentally. As a result, models trained on these datasets may struggle with rare, disordered, or membrane-bound proteins, which are underrepresented in existing databases.

Additionally, protein sequence diversity remains a challenge. Many novel proteins lack homologous sequences, making it difficult for evolutionary-based models to generate accurate predictions. While multiple sequence alignments (MSA) improve accuracy by leveraging evolutionary relationships, they are ineffective for proteins with no known relatives. In such cases, alternative strategies like self-supervised learning and generative models are being explored to predict structures without relying on evolutionary data.

Another major limitation is lack of experimental validation. Computationally predicted structures often require wet-lab verification to confirm their accuracy, but high-throughput validation techniques are still underdeveloped. Integrating computational predictions with experimental feedback loops could help refine model outputs and improve real-world applicability.

Model Interpretability and Black-Box Nature of Deep Learning: Many deep learning models used in protein structure prediction, particularly transformers and convolutional neural networks (CNNs), function as black-box models—meaning their decision-making processes are not fully interpretable. While models like AlphaFold2 achieve high accuracy, it remains unclear how specific residues influence the overall folding process.

Recent efforts in interpretable AI techniques focus on understanding these models. Methods such as attention visualization, feature attribution mapping, and residue importance scoring are being explored to trace model decisions. By improving interpretability, researchers can build more reliable and biologically meaningful models for protein structure prediction.

Predicting Side-Chain Conformations with High Precision: While deep learning models effectively predict protein backbone structures, accurately positioning side chains remains a challenge. Side-chain conformations play a crucial role in protein-protein interactions, enzymatic functions, and drug binding.

Traditional side-chain prediction methods rely on rotamer libraries—precomputed sets of likely conformations for each amino acid. However, these approaches lack precision in highly flexible regions. Hybrid models that integrate graph neural networks (GNNs) and physics-based energy calculations are being developed to improve side-chain accuracy by considering atomic-level interactions and steric clashes.

Multi-Protein Interactions and Quaternary Structure Prediction: Most deep learning models, including AlphaFold2, are optimized for single-protein structure prediction. However, proteins rarely function in isolation; they interact with other biomolecules to form quaternary structures and functional complexes.

For example, antibodies, enzymes, and receptor proteins require detailed modeling of protein-protein docking and ligand binding. Current tools such as AlphaFold-Multimer attempt to address this issue, but accurately modeling large-scale molecular assemblies remains an ongoing challenge. Developing multi-scale AI models that incorporate protein dynamics, ligand interactions, and post-translational modifications could enhance predictions in this area.

Energy-Based Refinements and Molecular Dynamics Simulations: Predicted protein structures often require post-processing and refinement to ensure structural stability and minimize steric clashes. Energy-based approaches, such as molecular dynamics (MD) simulations, provide a way to refine structures by simulating atomic movements over time.

Traditional MD tools, including AMBER and GROMACS, simulate the physical interactions of proteins in a virtual environment, allowing researchers to optimize folding pathways and identify energetically favorable conformations. However, these simulations are computationally expensive and require significant processing power.

To overcome this limitation, hybrid models that integrate deep learning with MD refinement are being developed. For instance, AlphaFold can generate an initial structure, which is then refined using MD simulations to achieve a more biologically relevant conformation. This combination allows for faster and more accurate predictions, reducing reliance on purely experimental techniques.

Computational Costs and Resource Limitations: Training large-scale deep learning models for protein structure prediction requires substantial computational resources, making these methods inaccessible to many researchers. Efforts are being made to develop lightweight and efficient models that require less computational power while maintaining accuracy. For example, AlphaFold-lite and low-memory deep learning architectures are being explored to make protein structure prediction more accessible to a wider scientific community.

4. Conclusion And Future Scope

The prediction of protein structures has evolved significantly over the years, transitioning from physics-based simulations and evolutionary modeling to advanced deep learning approaches.

Computational methods, particularly deep learning-driven models such as AlphaFold2 and RoseTTAFold, have demonstrated remarkable accuracy in predicting protein folding patterns. These advancements have bridged the gap between traditional experimental techniques like X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy, offering faster and cost-effective alternatives. However, several challenges remain, particularly in multi-protein interactions, modeling flexible regions, and improving side-chain accuracy.

Despite these breakthroughs, protein structure prediction still faces limitations in dataset biases, model interpretability, computational efficiency, and integration with real-world experimental validation. Many deep learning models depend heavily on high-quality sequence alignments, making it difficult to predict structures for novel or orphan proteins with no homologous sequences. Furthermore, while existing models can predict static structures with high precision, accurately simulating dynamic protein behaviors, ligand interactions, and post-translational modifications remains an ongoing challenge.

The future of protein structure prediction lies at the intersection of AI, quantum computing, and experimental biology. As computational models continue to evolve, their integration with real-world biochemical applications will pave the way for groundbreaking discoveries in drug development, synthetic biology, and personalized medicine. Addressing the existing challenges will not only refine our understanding of protein folding mechanisms but also unlock new possibilities for engineering novel biomolecules with tailored functions. By leveraging interdisciplinary innovations, protein modeling can become a cornerstone of next-generation biomedical research and therapeutic advancements.

References

[1] Jadhav, Swati and Vyavahare, Arati J. and Sharma, Manish, A Systematic Review on Protein Structure Prediction – Conventional and AI Methods (August 15, 2023). Available at SSRN: https://ssrn.com/abstract=4541252 or http://dx.doi.org/10.2139/ssrn.4541252

[2] Yang, Z., Zeng, X., Zhao, Y. et al. AlphaFold2 and its applications in the fields of biology and medicine. Sig Transduct Target Ther 8, 115 (2023). https://doi.org/10.1038/s41392-023-01381-z

[3] Ebrahimi, S.B., Samanta, D. Engineering protein-based therapeutics through structural and chemical design. Nat Commun 14, 2411 (2023). https://doi.org/10.1038/s41467-023-38039-x

[4] Gauto, D.F., Estrozi, L.F., Schwieters, C.D. et al. Integrated NMR and cryo-EM atomic-resolution structure determination of a half-megadalton enzyme complex. Nat Commun 10, 2697 (2019). https://doi.org/10.1038/s41467-019-10490-9

[5] Wang HW, Wang JW. How cryo-electron microscopy and X-ray crystallography complement each other. Protein Sci. 2017 Jan;26(1):32-39. doi: 10.1002/pro.3022. Epub 2016 Sep 7. PMID: 27543495; PMCID: PMC5192981.

[6] Heo L, Janson G, Feig M. Physics-based protein structure refinement in the era of artificial intelligence. Proteins. 2021 Dec;89(12):1870-1887. doi: 10.1002/prot.26161. Epub 2021 Jun 29. PMID: 34156124; PMCID: PMC8616793.

[7] Bouatta N, Sorger P, AlQuraishi M. Protein structure prediction by AlphaFold2: are attention and symmetries all you need? Acta Crystallogr D Struct Biol. 2021 Aug 1;77(Pt 8):982-991.

doi: 10.1107/S2059798321007531. Epub 2021 Jul 29. PMID: 34342271; PMCID: PMC8329862.

[8] Choopanian, P., Andressoo, JO. & Mirzaie, M. A fast approach for structural and evolutionary analysis based on energetic profile protein comparison. Nat Commun 16, 2231 (2025). https://doi.org/10.1038/s41467-025-57374-9

[9] Chen L, Li Q, Nasif KFA, Xie Y, Deng B, Niu S, Pouriyeh S, Dai Z, Chen J, Xie CY. AI-Driven Deep Learning Techniques in Protein Structure Prediction. Int J Mol Sci. 2024 Aug 1;25(15):8426. doi: 10.3390/ijms25158426. PMID: 39125995; PMCID: PMC11313475.

[10] Fu, Lihao& Gao, Yuan & Chen, Yongcan& Wang, Yanjing & Fang, Xiaoting& Tian, Shujun& Dong, Hao & Zhang, Yijian& Chen, Zichuan& Wang, Zechen& Hu, Shantong& Yi, Xiao & Si, Tong. (2024). Critical Assessment of Protein Engineering (CAPE): A Student Challenge on the Cloud. ACS synthetic biology. 13. 10.1021/acssynbio.4c00588.

[11] Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins. 2021 Dec;89(12):1607-1617. doi: 10.1002/prot.26237. Epub 2021 Oct 7. PMID: 34533838; PMCID: PMC8726744.

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

[13] Pereira, Joana, Adam J. Simpkin, Marcus D. Hartmann, et al., High-accuracy protein structure prediction in CASP14. Proteins: Structure, Function, and Bioinformatics2021;89(12):1687-1699.

[14] Malhotra, Sidharth & Walters, Robin. (2022). Secondary Protein Structure Prediction Using Neural Networks. 10.48550/arXiv.2208.11248.

[15] A. Rizqiana and Afiahayati, "Protein Secondary Structure Prediction using N-Grams and 1-Dimensional Convolutional Neural Network," 2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS), Himeji, Japan, 2024, pp. 1-6, doi: 10.1109/SCISISIS61014.2024.10759947.

[16] M. AlQuraishi, “End-to-end differentiable learning of protein structure,” bioRxiv. bioRxiv, p. 265231, 14-Feb-2018.

[17] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, “SchNet - a deep learning architecture for molecules and materials,” J. Chem. Phys., vol. 148, no. 24, Dec. 2017.

[18] B. Nebgen et al., “Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks,” J. Chem. Theory Comput., vol. 14, no. 9, pp. 4687–4698, Sep. 2018

[19] J. M. Jumper, N. F. Faruk, K. F. Freed, and T. R. Sosnick, “Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics,” PLoSComput. Biol., vol. 14, no. 12, p. e1006342, Dec. 2018.

[20] Wang, C., Wei, Y., Zhang, H. et al. Constructing effective energy functions for protein structure prediction through broadening attraction-basin and reverse Monte Carlo sampling. BMC Bioinformatics 20 (Suppl 3), 135 (2019). https://doi.org/10.1186/s12859-019-2652-5

[21] S. T. John and G. Csányi, “Many-Body Coarse-Grained Interactions Using Gaussian Approximation Potentials,” J. Phys. Chem. B, vol. 121, no. 48, pp. 10934–10949, Dec. 2017.

[22] M. K. Scherer, B. E. Husic, M. Hoffmann, F. Paul, H. Wu, and F. Noé, "Variational selection of features for molecular kinetics," J. Chem. Phys., vol. 150, no. 19, p. 194108, May 2019.
[23] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, "Deep generative models of genetic variation capture the effects of mutations," Nat. Methods, vol. 15, no. 10, pp. 816–822, Oct. 2018.
[24] R. Rao et al., "MSA Transformer," bioRxiv, p. 2021.02.12.430858, Feb. 2021.
[25] M. AlQuraishi, "Parallelized Natural Extension Reference Frame: Parallelized Conversion from Internal to Cartesian Coordinates," J. Comput. Chem., vol. 40, no. 7, pp. 885–892, Mar. 2019.
[26] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet - a deep learning architecture for molecules and materials," J. Chem. Phys., vol. 148, no. 24, Dec. 2017.
[27] B. Nebgen et al., "Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks," J. Chem. Theory Comput., vol. 14, no. 9, pp. 4687–4698, Sep. 2018.
[28] howdhury, R., Bouatta, N., Biswas, S. et al. Single-sequence protein structure prediction using a language model and deep learning. Nat Biotechnol 40, 1617–1623 (2022). https://doi.org/10.1038/s41587-022-01432-w
[29] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, Michal Linial, ProteinBERT: a universal deep-learning model of protein sequence and function, Bioinformatics, Volume 38, Issue 8, March 2022, Pages 2102–2110, https://doi.org/10.1093/bioinformatics/btac020
[30] Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, Wicky BIM, Courbet A, de Haas RJ, Bethel N, Leung PJY, Huddy TF, Pellock S, Tischer D, Chan F, Koepnick B, Nguyen H, Kang A, Sankaran B, Bera AK, King NP, Baker D. Robust deep learning-based protein sequence design using ProteinMPNN. Science. 2022 Oct 7;378(6615):49-56. doi: 10.1126/science.add2187. Epub 2022 Sep 15. PMID: 36108050; PMCID: PMC9997061.
[31] D'Souza, S., Prema, K.V., Balaji, S. et al. Deep Learning-Based Modeling of Drug–Target Interaction Prediction Incorporating Binding Site Information of Proteins. Interdiscip Sci Comput Life Sci 15, 306–315 (2023). https://doi.org/10.1007/s12539-023-00557-z
[32] Kiouri, D. P., Batsis, G. C., & Chasapis, C. T. (2025). Structure-Based Approaches for Protein–Protein Interaction Prediction Using Machine Learning and Deep Learning. Biomolecules, 15(1), 141. https://doi.org/10.3390/biom15010141
[33] Wei Zheng, QiqigeWuyun, Xiaogen Zhou, Yang Li, Lydia Freddolino, Yang Zhang, LOMETS3: integrating deep learning and profile alignment for advanced protein template recognition and function annotation, Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W454–W464, https://doi.org/10.1093/nar/gkac248
[34] Mirko Torrisi, Gianluca Pollastri, Brewery: deep learning and deeper profiles for the prediction of 1D protein structure annotations, Bioinformatics, Volume 36, Issue 12, June 2020, Pages 3897–3898, https://doi.org/10.1093/bioinformatics/btaa204
[35] Pearce R, Zhang Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. Curr Opin Struct Biol. 2021 Jun;68:194-207. doi: 10.1016/j.sbi.2021.01.007. Epub 2021 Feb 24. PMID: 33639355; PMCID: PMC8222070.

[36] S. I. Jalal, J. Zhong and S. Kumar, "Protein Secondary Structure Prediction using Multi-input Convolutional Neural Network," 2019 Southeast Con, Huntsville, AL, USA, 2019, pp. 1-5, doi: 10.1109/SoutheastCon42311.2019.9020333.secon

[37] Enireddy, V., Karthikeyan, C. & Babu, D.V. OneHotEncoding and LSTM-based deep learning models for protein secondary structure prediction. Soft Comput 26, 3825–3836 (2022). https://doi.org/10.1007/s00500-022-06783-9

[38] Dong B, Liu Z, Xu D, Hou C, Dong G, Zhang T, Wang G. SERT-StructNet: Protein secondary structure prediction method based on multi-factor hybrid deep model. Comput Struct Biotechnol J. 2024 Mar 22;23:1364-1375. doi: 10.1016/j.csbj.2024.03.018. PMID: 38596312; PMCID: PMC11001767.

[39] M. McPartlon, & J. Xu, An end-to-end deep learning method for protein side-chain packing and inverse folding, Proc. Natl. Acad. Sci. U.S.A. 120 (23) e2216438120, https://doi.org/10.1073/pnas.2216438120 (2023).

[40] Liu J, Zhang C, Lai L. GeoPacker: A novel deep learning framework for protein side-chain modeling. Protein Science. 2022; 31(12):e4484. https://doi.org/10.1002/pro.4484

[41] McPartlon M, Xu J. An end-to-end deep learning method for protein side-chain packing and inverse folding. Proc Natl Acad Sci U S A. 2023 Jun 6;120(23):e2216438120. doi: 10.1073/pnas.2216438120. Epub 2023 May 30. PMID: 37253017; PMCID: PMC10266014.

[42] Audagnotto, M., Czechtizky, W., De Maria, L. et al. Machine learning/molecular dynamic protein structure prediction approach to investigate the protein conformational ensemble. Sci Rep 12, 10018 (2022). https://doi.org/10.1038/s41598-022-13714-z

[43] Chen, L., Li, Q., Nasif, K. F. A., Xie, Y., Deng, B., Niu, S., Pouriyeh, S., Dai, Z., Chen, J., & Xie, C. Y. (2024). AI-Driven Deep Learning Techniques in Protein Structure Prediction. International Journal of Molecular Sciences, 25(15), 8426. https://doi.org/10.3390/ijms25158426

[44] Adhikari, Badri. (2019). DEEPCON: Protein Contact Prediction using Dilated Convolutional Neural Networks with Dropout. Bioinformatics (Oxford, England). 36. 10.1093/bioinformatics/btz593.

[45] Lee M. Recent Advances in Deep Learning for Protein-Protein Interaction Analysis: A Comprehensive Review. Molecules. 2023 Jul 2;28(13):5169. doi: 10.3390/molecules28135169. PMID: 37446831; PMCID: PMC10343845.

[46] Alakus T.B., Turkoglu I. A Novel Protein Mapping Method for Predicting the Protein Interactions in COVID-19 Disease by Deep Learning. Interdiscip. Sci. Comput. Life Sci. 2021;13:44–60. doi: 10.1007/s12539-020-00405-4.

[47] Pakhrin, S. C., Shrestha, B., Adhikari, B., & KC, D. B. (2021). Deep Learning-Based Advances in Protein Structure Prediction. International Journal of Molecular Sciences, 22(11), 5553. https://doi.org/10.3390/ijms22115553

[48] Si, D.; Moritz, S.A.; Pfab, J.; Hou, J.; Cao, R.; Wang, L.; Wu, T.; Cheng, J. Deep Learning to Predict Protein Backbone Structure from High-Resolution Cryo-EM Density Maps. Sci. Rep. 2020, 10, 4282

[49] N. Hiranuma, H. Park, M. Baek, I. Anishchenko, J. Dauparas, and D. Baker, “Improved protein structure refinement guided by deep learning based accuracy estimation,” Nat. Commun., vol. 12, no. 1, p. 1340, Dec. 2021.
[50] P. Sturmfels, J. Vig, A. Madani, and N. F. Rajani, “Profile Prediction: An Alignment-Based Pre-Training Task for Protein Sequence Models,” arXiv, Nov. 2020.
[51] A. W. Senior et al., “Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13),” Proteins Struct. Funct. Bioinforma., vol. 87, no. 12, pp. 1141–1148, Dec. 2019.
[52] “AlphaFold: a solution to a 50-year-old grand challenge in biology, DeepMind.” [Online]. Available:https://deepmind.com/blog/article/alphafold-asolution-to-a-50-year-old-grand-challenge-inbiology. [Accessed: 07-Apr-2021].
[53] A. W. Senior et al., “Improved protein structure prediction using potentials from deep learning,” Nature, vol. 577, no. 7792, pp. 706–710, Jan. 2020.
[54] C. Mirabello and B. Wallner, “RAWMSA: End-toend Deep Learning using raw Multiple Sequence Alignments,” PLoS One, vol. 14, no. 8, Aug. 2019.
[55] S. M. Kandathil, J. G. Greener, A. M. Lau, and D. T. Jones, “Deep learning-based prediction of protein structure using learned representations of multiple sequence alignments,” bioRxiv. bioRxiv, 27-Nov-2020.
[56] J. Jumper et al., “High Accuracy Protein Structure Prediction Using Deep Learning,” 2020.
[57] A. W. Senior et al., “Improved protein structure prediction using potentials from deep learning,” Nature, vol. 577, no. 7792, pp. 706–710, Jan. 2020.
[58] [58] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, “Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model,” PLOS Comput. Biol., vol. 13, no. 1, p. e1005324, Jan. 2017.
[59] W. Russ et al., “Evolution-based design of chorismate mutase enzymes,” bioRxiv, p. 2020.04.01.020487, Apr. 2020.
[60] T. Blazejewski, H. I. Ho, and H. H. Wang, “Synthetic sequence entanglement augments stability and containment of genetic information in cells,” Science (80)., vol. 365, no. 6453, pp. 595-598, Aug. 2019.
[61] A. Rives et al., “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,” bioRxiv. bioRxiv, p. 622803, 29-Apr-2019.
[62] M. Heinzinger et al., “Modeling the language of life - Deep learning protein sequences,” bioRxiv. bioRxiv, p. 614313, 19-Apr-2019.
[63] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, “Transformer protein language models are unsupervised structure learners,” bioRxiv. p. 2020.12.15.422761, 15-Dec-2020.
[64] D. T. Jones and S. M. Kandathil, “High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features,” Bioinformatics, vol. 34, no. 19, pp. 3308–3315, Oct. 2018.
[65] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, “Improved protein structure prediction using predicted inter-residue orientations,” bioRxiv. bioRxiv, 18-Nov-2019.

[66] J. Ingraham, A. Riesselman, C. Sander, D. Marks, and H. M. School, “Learning Protein Structure with A Differentiable Simulator,” Sep. 2018.
[67] M. M. Mohamed Mufassirin, M. A. H. Newton, J. Rahman and A. Sattar, "Multi-S3P: Protein Secondary Structure Prediction With Specialized Multi-Network and Self-Attention-Based Deep Learning Model," in IEEE Access, vol. 11, pp. 57083-57096, 2023, doi: 10.11.
[68] A. Elnaggar et al., “ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Deep Learning and High-Performance Computing,” bioRxiv, Jul. 2020.
[69] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf., vol. 1, pp. 4171–4186, Oct. 2018.
[70] T. Bepler and B. Berger, “Learning protein sequence embeddings using information from structure,” arXiv, Feb. 2019.
[71] A. Rives et al., “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences,” bioRxiv. bioRxiv, p.622803, 29-Apr-2019.
[72] J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. F. Rajani, “BERTology Meets Biology: Interpreting Attention in Protein Language Models,” bioRxiv, Jun. 2020.
[73] R. Rao et al., “Evaluating Protein Transfer Learning with TAPE,” bioRxiv, Jun. 2019.
[74] P. Sturmfels, J. Vig, A. Madani, and N. F. Rajani, “Profile Prediction: An Alignment-Based Pre-Training Task for Protein Sequence Models,” arXiv, Nov. 2020.
[75] T. Sercu et al., “Neural Potts Model” Open Review, 2020, pp. 1-13.
[76] A. Madani et al., “ProGen: Language modeling for protein generation,” bioRxiv. bioRxiv, p.2020.03.07.982272, 08-Mar-2020
[77] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, “Less is more: Sampling chemical space with active learning,” J. Chem. Phys., vol. 148, no. 24, p. 241733, Jun. 2018.
[78] S. Raman et al., “Structure prediction for CASP8 with all-atom refinement using Rosetta,” Proteins Struct. Funct. Bioinforma., vol. 77, no. SUPPL. 9, pp. 89-99, 2009
[79] Wenhao Gao, Sai Pooja Mahajan, Jeremias Sulam, Jeffrey J. Gray (2020). Deep Learning in Protein Structural Modeling and Design, Pattrns, Volume 1, Issue 9, 11 December 2020, 100142. https://doi.org/10.1016/j.patter.2020.100142
[80] [80] Jang, Woo Dae, Gi Bae Kim, Yeji Kim and Sang Yup Lee. Applications of artificial intelligence to enzyme and pathway design for metabolic engineering. Current Opinion in Biotechnology 2022;73:101-107
[81] AlQuraishi M. Machine learning in protein structure prediction. Curr Opin Chem Biol. 2021 Dec;65:1-8. doi: 10.1016/j.cbpa.2021.04.005. Epub 2021 May 18. PMID: 34015749.
[82] Al Quraishi, Mohammed. Machine learning in protein structure prediction. Current Opinion in Chemical Biology 2021;65:1-8.
[83] Yuvaraj, Natarajan, Kannan Srihari, SelvarajChandragandhi, RajanArshath Raja, Gaurav Dhiman and Amandeep Kaur. Analysis of protein-ligand interactions of SARS-Cov-2 against selective drug-using deep neural networks. Big Data Mining and Analytics2021;4(2):76-83.
[84] Akhter, Nasrin, Kazi Lutful Kabir, Gopinath Chennupati, Raviteja Vangara, Boian Alexandrov, Hristo N. Djidjev and Amarda Shehu. Improved protein decoy selection via non-negative

matrix factorization. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2021.

[85] Burley, StephenK and Helen M. Berman. Open-access data: A cornerstone for artificial intelligence approaches to protein structure prediction. Structure 2021.

[86] Bryant, P., Pozzati, G., and Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. Nat. Commun. 13, 1265. doi:10.1038/s41467-022-28865-w

[87] Drake, Z.C., Seffernick, J.T. & Lindert, S. Protein complex prediction using Rosetta, AlphaFold, and mass spectrometry covalent labeling. Nat Commun 13, 7846 (2022). https://doi.org/10.1038/s41467-022-35593-.