Review Article

BIOMEDICAL Journal of Scientific & Technical Research

ISSN: 2574 -1241

DOI: 10.26717/BJSTR.2026.64.010069

# Predicting the 3D Structure of Proteins Using AI Tools: A Review

**Theodoros Christodoulou***

*School of Medicine, Democritus University of Thrace, Alexandroupolis 68100, Greece*

***Corresponding author:** Theodoros Christodoulou, School of Medicine, Democritus University of Thrace, Alexandroupolis 68100, Greece

**ARTICLE INFO**

**Received:** January 19, 2026

**Published:** February 02, 2026

**Citation:** Theodoros Christodoulou. Predicting the 3D Structure of Proteins Using AI Tools: A Review. Biomed J Sci & Tech Res 64(4)-2026. BJSTR. MS.ID.010069.

**ABSTRACT**

The protein-folding problem has long been one of the most significant challenges in molecular biology, due to the intricate complexity of protein structures, the mechanisms underlying the folding process, and the high costs and time-consuming nature of experimental techniques for determining atomic positions within a molecule. Recent advances in computational power and artificial intelligence (AI) have significantly advanced the ability to address this problem, enabling rapid, cost-effective, and highly accurate analyses of complex biological phenomena. AlphaFold 2, an AI program developed by DeepMind (Google), has demonstrated the ability to predict the 3D structures of nearly all proteins in the Protein Data Bank with an accuracy approaching that of experimental methods. This paper provides a concise review of AlphaFold and other AI-based methods for predicting tertiary protein structures, highlighting their real-world applications, impacts, limitations, and the ongoing challenges that remain to be addressed in this evolving field.

**Keywords:** Protein Folding; Amino Acids; Artificial Intelligence; Deep Learning; AlphaFold

**Abbreviations:** PTMs: Post-Translational Modifications; PDB: Protein Data Bank; CASP: Critical Assessment of Techniques for Protein Structure Prediction; PFP: Protein-Folding Problem; NMR: Nuclear Magnetic Resonance; MSA: Multiple Sequence Alignment

## Introduction

The human body contains approximately 20,000 protein-coding genes within the human genome [1]. These genes are responsible for producing hundreds of thousands of distinct proteins and their variants, collectively called the human proteome. A single gene can generate multiple proteins through processes such as alternative splicing (AS) and post-translational modifications (PTMs), which give rise to different protoforms (protein species) [2]. The total proreform space is estimated to range from 1 to 10 million, although it is not yet fully characterized. This immense diversity explains why proteins constitute the majority (50%) of the dry weight of a human cell, while the remaining dry weight is composed of lipids (20%), carbohydrates (10%), nucleic acids (10%), inorganic ions, and other small molecules. When considering all organisms, the total number of known proteins continues to grow rapidly, with current estimates suggesting approximately 200 million proteins [3]. To manage the vast amount of information generated by such an extensive dataset, specialized databases have been developed to catalog protein structures, sequences, interactions, and functionalities. Prominent examples include UniProt, a comprehensive resource for protein sequence and functional information, the Protein Data Bank (PDB), which archives 3D structural data of biomolecules, including proteins and nucleic acids and InterPro, a database that classifies proteins into families and predicts functional domains and important sites. These databases collectively enable researchers to study protein behavior, interactions, and evolution on a global scale.

Despite the vast complexity and diversity of the protein space, the functionality of each individual protein is unique and essential for maintaining the overall quality, functionality, and phenotypic characteristics of an organism. During the intricate process of protein folding, even a minor alteration at any step can result in a protein with drastically different functionality, which may be either beneficial or detrimental to the organism. A well-known example is the tumor suppressor protein p53, which plays a critical role in preventing cancer. A single-point mutation, such as the R175H mutation-where arginine is substituted with histidine - causes the protein to misfold. Conse-

quently, the misfolded p53 loses its ability to regulate the cell cycle and repair damaged DNA, ultimately promoting tumor growth and metastasis [4]. For the reasons outlined above, understanding the mechanisms underlying the well-known "protein-folding problem" and developing frameworks to predict the three-dimensional structures of proteins are critical to advancing our knowledge of human biology. Decoding a protein's shape and molecular structure has broad practical applications, ranging from drug discovery to the design of proteins with specific properties. The protein-folding problem was first introduced by Anfinsen in 1961 [5], who proposed that a protein's primary goal is to minimize the free energy resulting from the interactions between the amino acids in its sequence. Since then, numerous researchers have worked on computational methods and algorithms to predict the 3D structure of a protein based on its primary sequence of amino acids. In 1994, the Critical Assessment of Techniques for Protein Structure Prediction (CASP) competition was established, held every two years, where teams tested their artificial intelligence (AI) models using target protein sequences with known experimental structures that were not yet included in the Protein Data Bank (PDB). During the CASP14 competition in 2020, DeepMind's AlphaFold2 achieved a breakthrough by accurately predicting protein folding, reaching levels of accuracy comparable to experimental methods. For this achievement, Demis Hassabis, the company's director, and lead researcher John Jumper were awarded a prize of $3 million.

In the following section, the problem of predicting the three-dimensional structure of proteins will be discussed, alongside the key artificial intelligence (AI) and machine learning (ML) techniques explored in the literature to address this challenge, with particular focus on the AlphaFold2 model. A concise literature review of some popular studies on this topic will be presented in Chapter 3, utilizing PubMed search and Google Scholar to gather relevant sources. On Chapter 4, real-world applications and implications of solving this problem will be explored, along with an examination of the limitations and potential avenues for future research in this field.

## Protein Folding and AlphaFold2 Methodology

Proteins are polypeptide chains made up of amino acids, and in order to perform their biological functions, most of them must adopt a unique three-dimensional shape. This process is known as protein folding [6]. Initially, a protein exists in its primary structure, which refers to its amino acid sequence. It then progresses to the secondary structure, where localized regions of the polypeptide backbone adopt specific conformations, such as alpha helices and beta sheets. The protein continues folding into its tertiary structure, involving further interactions among the R groups of the amino acids. Some proteins even reach a quaternary structure, where multiple polypeptide subunits come together. Predicting a protein's final 3D structure based solely on its primary sequence is an extremely challenging task. According to Levinthal's paradox [7], the number of possible conformations for a protein can be as large as 100100, suggesting that the protein would take years to fold into its final state. However, in reality, this process occurs in a matter of minutes or even seconds.

As noted in [8], the protein-folding problem (PFP) can be divided into three distinct problems that can be addressed individually: the folding code (the interatomic forces in the primary structure that ultimately determine the fully folded protein), the kinetic aspects (the speed at which a protein folds), and the computational problem (using in silico and AI-based approaches to predict a protein's tertiary structure based solely on its primary sequence). The first two problems fall under the broader category of physical interactions between the amino acids in the polypeptide chain, while the third is a bioinformatics challenge, closely linked to evolutionary history. Experimental techniques such as nuclear magnetic resonance (NMR), X-ray crystallography, and cryo-electron microscopy can be employed to predict the 3D structure of proteins. However, these experimental methods are both expensive and time-consuming, and to date, they have only been able to solve approximately 200,000 protein structures stored in the PDB [9]. As a result, research in this field has largely shifted towards AI-driven approaches. The algorithms used in this domain can be

Broadly classified into three categories:

1. Homology or template-based modeling,
2. De novo modeling, and
3. Machine learning (ML)-based modeling. The first category involves models that rely on sequence alignment techniques (such as multiple sequence alignment, MSA) to search databases like PDB and UniProt for similar proteins (templates) based on evolutionary history. The second category, de novo modeling, does not rely on templates but instead uses the laws of physics to explore possible conformations and select the one with the lowest free energy. The third category, ML-based modeling, primarily uses deep learning methods, including neural networks, to predict the 3D structure of target proteins based on known structures. This short review focuses on this latter category, highlighting AI-developed models like AlphaFold, which have made significant contributions by providing valuable insights and computationally solving many protein structures.

Among the AI models presented in the CASP contests the last few years, AlphaFold2 [10] was the first computational method to predict the protein structures with atomic accuracy even with protein with not known homologs, with an accuracy really close to that of the experimental results. The main goal of AlphaFold is to predict the coordinates of the heavy atoms in the final folded structure, given as input specific amino acid sequence and searching for homologs in databases with specific algorithms, through the MSA process. Its architecture is actually a system of multiple neural network components that work iteratively into a pipeline, to refine the predictions of the final 3d structure.

As a first step, the primary amino acid sequence and its aligned sequences are input into the model, generating both the multiple sequence alignment (MSA) feature map between the target protein and its homologs, and the Pairwise Representation (PR) table of the amino acids. These matrices are processed by 48 repeated layers of a novel neural network block called the Evo former. Each block has an MSA representation that updates the PR matrix, using attention mechanisms along both rows and columns to focus on regions of the protein that are crucial for folding. The transition layer integrates sequence information with evolutionary context, and the final MSA representation updates the PR matrix through an element-wise outer product mean.

The PR matrix contains additional components, including triangle multiplicative updates and triangle self-attention mechanisms, which ensure that amino acids interact in a way that produces a physically valid 3D structure. The outputs of this stage-the final pairwise and single representations-encapsulate spatial and evolutionary information that is then used by the Structure Model module to predict the protein's 3D conformation. This module comprises eight blocks, where an initial guess of the 3D coordinates of the heavy atoms is progressively refined. The Cα atoms are positioned first in compliance with the global frame, followed by rotations and translations of each amino acid to achieve the correct spatial arrangement. Side chains are placed only after the backbone is set, with the Invalid Point Attention (IPA) mechanism ensuring that these transformations respect local 3D geometry constraints. Additional updates guarantee consistency between each residue's local frame and the overall protein geometry.

Finally, side chain x-angle rotations are computed, along with confidence metrics such as pLDDT (predicted local distance difference test) and pTM (predicted template modeling score), which indicate how closely the predicted structure resembles the real protein. The Final Average Distance Error (FAPE) compares predicted atom positions with experimentally confirmed positions, producing the final 3D shape. To enhance accuracy, the entire pipeline-including Evoformer and Structure Model components-is recycled three times, using the current predicted structure as the new initial input. This iterative refinement adjusts atomic positions as needed, resulting in a highly accurate prediction of the protein's structure. The basic components of AF2 pipeline are shown graphically in Figure 1. The biological data that was used to train AlphaFold2 are all the protein sequences stored in the PDB that were experimentally defined. In addition to this, self-distillation was also applied to unlabeled proteins that were predicted from AlphaFold2 itself. Specifically, 75% of the dataset was protein structures from PDB and the rest 25% was from the new dataset predicted by the model. This method enhances the accuracy of the model to a great extent.

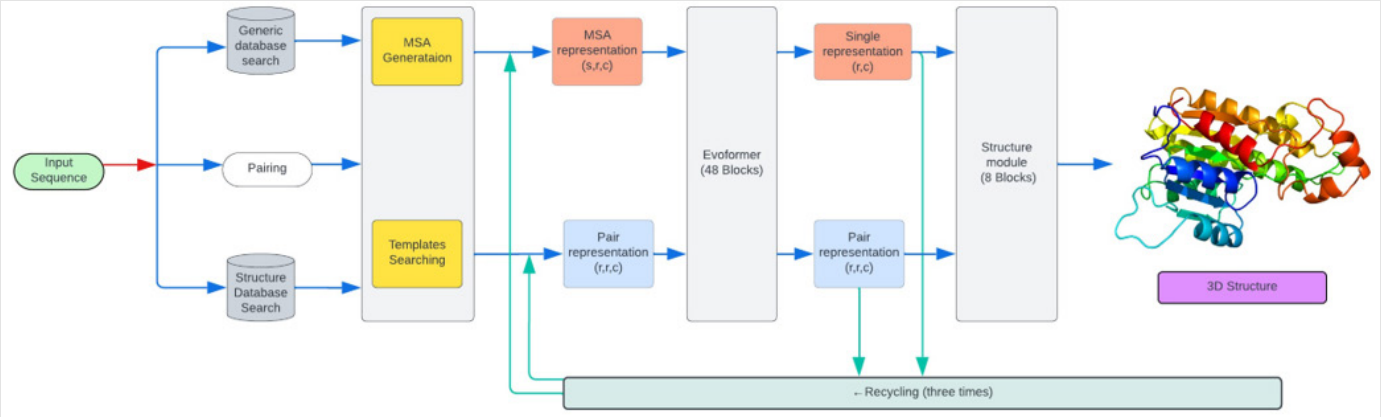


**Figure 1:** AlphaFold2 Neural Network Pipeline.

## Literature Review

In the field of protein folding, AlphaFold has established itself as the state-of-the-art model. On May 8, 2024, DeepMind introduced the latest version, AlphaFold3 [11], which marks a significant advancement in the prediction of protein structures. Beyond achieving higher accuracy in structural predictions, AlphaFold3 extends its capabilities to encompass a broader spectrum of biomolecules, including nucleic acids, ions, ligands, and protein-protein complexes. Key components of the model have been revamped, with the Evoformer and Structure modules being replaced by the Pairformer and Diffusion model, respectively. Despite these changes, the foundational neural network architecture, including attention mechanisms, remains intact. The training process for AlphaFold3 leverages data from the Protein Data Bank (PDB), applying tailored filtering for different biomolecular categories, such as ligands, ions, proteins, and complexes. The accuracy of its predictions is rigorously evaluated using specialized metrics, including DockQ, LDDT, and pocket-aligned root mean square deviation (RMSD).

In recent years, deep learning (DL) and ML approaches have revolutionized the field of protein structure prediction, particularly through the development of AI models that have gained prominence in CASP competitions. A comprehensive review by Chen, et al. [12] highlights several state-of-the-art deep learning models, including AlphaFold (versions 1, 2, and 3), RoseTTAFold, ProteinBERT, DeepFold, OmegaFold, ESMFold, Swiss-Model, Rosetta, and I-TASSER. These models represent the forefront of AI-driven protein folding, showcasing the diversity and sophistication of techniques now available. Chen et al. also include a thorough discussion of the evaluation metrics commonly used to assess the performance of these models. These include the Global Distance Test Total Score (GDT TS), Template Modeling Score (TM-score), Z-score (the deviation of a model's GDT TS score from the mean GDT TS score of all models), and LDDT. The majority of deep learning-based models rely on sophisticated encoders, such as one-hot encoding or Position Specific Scoring Matrices (PSSM), to transform protein sequences (composed of amino acid residues) into numerical vectors. These encoded sequences enable the models to extract meaningful features, significantly improving the accuracy of structure prediction.

## AI Models for Protein Structure Prediction

RoseTTAFold [13] shares many similarities with AlphaFold but is designed as a more streamlined and com- putationally efficient model. Like AlphaFold, RoseTTAFold leverages Multiple Sequence Alignments (MSA) and attention mechanisms within its neural network framework. However, it operates on a unique three-track system: The first track processes the amino acid sequence of the target protein, the second track extracts co-evolutionary information from the MSA map, and the third track generates the 3D structure, which is then iteratively refined using graph-based deep learning techniques (Graph Neural Networks, GNN). This structure allows Rosetta Fold to be significantly faster and less resource-intensive than AlphaFold, making it more accessible and user-friendly, particularly for high-throughput applications. Deep Fold [14], while similar in approach to the aforementioned models, incorporates Convolutional Neural Networks (CNN), attention mechanisms, MSA, and GNN techniques within its architecture. The model consists of several key components, including DeepMSA2, which constructs the MSA matrix by querying multiple sequence databases, and Deep Potential, which generates distance and contact maps as well as inter- residue orientations. Additionally, DeepFold employs L-BFGS folding simulations to refine and generate the final 3D structure. Unlike other models, Deep Fold does not directly predict the 3D structure but instead excels in protein-protein interaction predictions. While it is less efficient and slower in structure prediction compared to models like AlphaFold and Rosetta Fold, Deep Fold's strength lies in its ability to model protein interactions with higher accuracy.

Omega Fold [15] introduces a novel approach to protein structure prediction by diverging from the reliance on MSA and homologous modeling. Instead, Omega Fold predicts protein structures using only the single primary amino acid sequence, making it both simpler and faster. This is particularly advantageous for predicting the structures of proteins without a family tree that lack closely related homologs, or for rapidly evolving proteins such as antibodies. The core of Omega Fold consists of two key components: the Protein Language Model (Omega PLM), which is responsible for sequence modeling, and the Geoformer, a neural network comprising 50 blocks that handles structure prediction. The final stage involves a structure refinement step, using a model with 8 blocks, like AlphaFold, to predict the coordinates of the heavy atoms in the protein. Esmond [16], developed by Meta, leverages a custom protein language model called ESM-2, which draws inspiration from transformer architectures used in natural language processing (NLP), such as GPT and BERT. Unlike the previous models, ESM-Fold directly predicts a protein's 3D structure from its amino acid sequence, making it particularly useful for proteins with few known homologs. The model's simplicity and streamlined approach contribute to its faster performance, particularly for large-scale predictions.

Overall, ESMFold is primarily based on a transformer-based architecture and does not incorporate as many neural network components as other models, which further enhances its computational efficiency. ProteinBERT [17] represents a more generalized AI solution, pretrained on an extensive dataset of 106 million protein sequences derived from the Uniport and UniRef90 databases. Unlike models focused on 3D structure prediction, Protein BERT specializes in a variety of bioinformatics tasks, including protein classification, functional prediction, feature embedding, and Gene Ontology (GO) annotation. While it does not predict protein structures, it leverages foundational deep learning techniques, such as transformers and global attention mechanisms, to effectively capture complex sequence patterns and relationships within large- scale protein data.

## AI Integration Platforms

In addition to standalone AI models for protein structure prediction, several platforms integrate AI methods to enhance their capabilities. For instance, Swiss-Model [18] is a user-friendly, web-based platform that predicts 3D protein structures based on template modeling. It searches databases such as the AlphaFold DB to identify potential templates for a target protein when no experimental structure is available. Additionally, Swiss-Model can predict quaternary structures and employs machine learning techniques, such as Support Vector Machines (SVMs), to account for evolutionary constraints among homologs. Users input a single amino acid sequence in FASTA format, and the platform generates a 3D model of the target protein. In contrast, Rosetta application [19] rely on ab initio (de novo) methods to predict protein structures. Unlike template-based approaches, Rosetta is more efficient in designing entirely new protein structures with specific functionalities, as it does not rely on homologs or structural templates. A notable feature of Rosetta is its energy minimization pro-

cess, which refines the atomic positions in 3D structures to improve stability and accuracy.

This refinement capability extends to both computationally predicted and experimentally determined structures, further enhancing their quality and reliability. Another platform, I-TASSER [20], combines threading, fragment-based assembly, and structure refinement in its protein modeling pipeline. The threading process searches the PDB for homologs and templates, utilizing machine learning algorithms to uncover sequence-structure relationships. Subsequently, fragment assembly algorithms organize and classify individual protein fragments to construct a preliminary structure. The final structure is refined through Monte Carlo simulations, adhering to biochemical principles and appropriate constraints. Importantly, when no suitable templates are available, I-TASSER employs ab initio methods to generate models, making it a versatile tool for a wide range of protein structure prediction scenarios.

(Table 1) summarizes some of the main characteristics of all the models described in sections 3.1 and 3.2.

**Table 1:** Main characteristics of protein-folding AI models.

| Model | Key Compo- nents | Features | Strengths | Limitations |
|---|---|---|---|---|
| AlphaFold3 [11] | Pairformer, Diffusion model, Neural Networks | High accuracy; sup- ports proteins, nu- cleic acids, ligands, ions, complexes | Best accuracy; broad biomolecule scope | Computationally expensive |
| RoseTTAFold [13] | MSA, Attention mecha- nisms, 3-track system | Faster, efficient 3D struc- ture prediction | Faster than Al phaFold | Slightly less accurate |
| DeepFold [14] | CNN, MSA, GNN, L-BFGS | Focuses on protein interactions, not direct structure | Strong in protein inter- actions | Slower structure prediction |
| OmegaFold [15] | OmegaPLM, Geoformer | Predicts structures from amino acid sequence only | Faster, simpler, good for evolving proteins | Less accurate for some pro- teins |
| ESMFold [16] | ESM-2 (trans former), direct 3D structure pre- diction | Efficient, fast, for large- scale predic- tions | Fast performance | Simpler, less com- plex than others |
| ProteinBERT [17] | Transformer, pretrained On protein sequences | Sequence analysis, classification, functional prediction | Good for bioinformatics tasks | No 3D structure prediction |
| Swiss-Model [18] | Template modeling, SVMs | Web-based, template-based 3D structure predic- tion | User-friendly, tem- plate-based | Relies on avail- able templates |
| Rosetta [19] | Ab initio, Energy minimi- za- tion | De novo design, structure refinement | Flexible in struc- ture design | Computationally intensive |
| I-TASSER [20] | Threading, Fragment- based assem bly, Ab initio | Combines template- based and ab initio methods | Works without templates | Slower than templatebased methods |

## Applications

Rather than simply cataloging the most prominent AI models in the protein-folding field, it is equally critical to highlight their transformative applications in medicine and biology. AI-powered protein structure predic- tion models like AlphaFold have revolutionized various domains, including drug discovery, protein engineering (designing novel proteins), predicting protein-protein interactions, and analyzing misfolded proteins to better understand diseases such as cancer and neurodegenerative disorders. Beyond healthcare, these models also find applications in industries like food technology and manufacturing. In this paper, we briefly discuss some of the studies that illustrate the practical applications of these AI models in some of these fields. For example, Jimenez, et al. [21] employed CNNs and ML techniques to analyze 7,622 proteins from the scPDB database, identifying potential drug-binding sites. By conceptualizing protein structures as 3D images, their model, DeepSite, used computer vision approaches to predict binding sites. The evaluation metrics, such as the distance to the center of the binding site (DCC) and discretized volumetric overlap (DVO), demonstrated the model's efficacy, achieving an average DVO of 65%. Zhang, et al. [22] leveraged a molecular docking program called Glide to virtually screen the performance of 37 familiar drug targets, each represented by both exper- imental and AlphaFold 2 (AF2)-predicted structures. Their findings revealed that AF2-predicted structures performed comparably to experimental structures, with high enrichment factor values indicating the ability to identify active compounds binding to drug targets effectively. This approach facilitates the development of more targeted and efficient drugs for treating conditions such as cancer, neurodegenerative disorders, and infectious diseases [23]. Additionally, several studies have

focused on identifying small compounds targeting specific proteins. For instance, AI techniques have been applied to discover compounds targeting CDK20, WSB1, JMJD8, and other proteins, offering promising pathways for novel therapeutic strategies [24-26]. These applications exemplify the significant potential of AI in advancing both fundamental biological research and real-world medical innovations.

Designing new proteins has become a cornerstone of advancements in rapidly growing fields such as biotechnology, therapeutics, drug discovery, and bioengineering. The sheer vastness of the search space for all potential proteins makes it infeasible to experimentally solve every structure and uncover new biological functionalities with unknown properties. AI models like AlphaFold 2 (AF2) and others discussed earlier offer powerful tools for de novo protein design, enabling the creation of proteins with specific, desired functionalities even in the absence of known templates. In their comprehensive review, Pan, et al. [27] explore recent innovations in de novo protein design. One promising approach involves making targeted refinements or minor alterations to the backbone structure of proteins already stored in the PDB, rather than redesigning entire proteins from scratch. This method retains structural integrity while introducing novel functionalities. The subsequent step focuses on sampling and optimizing side chains, which play a crucial role in determining a protein's biochemical properties and functionality.

Given the vast combinatorial possibilities in side-chain sampling, machine learning algorithms and advanced simulation techniques are employed to narrow down the potential amino acid types, significantly reducing the search space and expediting the design process. Anishchenko, et al. [28] explored the generation of synthetic proteins by inputting manually generated random amino acid sequences into the trRosetta model to produce an initial residue-residue distance map. These sequences were then subjected to Monte Carlo simulations, which revealed that network-synthesized genes encoded 129 protein sequences. Upon inserting these synthetic genes into E. coli, 27 of the resulting proteins successfully folded into stable, monodisperse forms. Experimental validation confirmed that these artificially generated proteins closely resembled naturally occurring proteins, demonstrating the potential of AI-driven protein design. Another important expansion of protein-folding AI models, except from focusing only on sole target proteins, is that they are capable of predicting protein-protein interactions which is very usefully in discovering cellular pathways which can leads to many diseases such as cancer, metabolic or autoimmune diseases. These interactions can be between two or more proteins which are transferring "signals" through signaling proteins [29]. In their study, Evans, et al. [30] extended the AlphaFold2 model to predict multimeric protein complexes, introducing a new model named AlphaFold-Multimer. When tested on 17 heterodimer proteins, AlphaFold- Multimer achieved medium accuracy on approximately 13 targets and high accuracy on 7 targets, as assessed by the DockQ score. Additionally, the model was evaluated on a dataset of 4,000 protein complexes, achieving successful interface predictions for 67% of heteromeric and 69% of homomeric complexes, even in cases with no close structural templates. Following the success of AlphaFold-Multimer, further studies expanded its application, including the prediction of complexes such as PHF14-HMG20A and CYP102A1, showcasing the growing versatility of protein-folding AI models [31,32].

## Discussion

The advent of AI has significantly advanced our understanding of various biological challenges, particularly the protein folding problem. These breakthroughs, while transformative, have also highlighted the complexities of evaluating AI-driven models. Despite the considerable success of state-of-the-art models, they are not immune to errors and still possess inherent limitations. In the context of protein structure prediction, models are typically evaluated through competitions like CASP (Critical Assessment of Structure Prediction), where their performance is assessed using specialized statistical measures designed for this task. Additionally, AI models developed to assess the performance and accuracy of primary prediction models, such as ModFold8 [33], are also commonly used for evaluation purposes. However, directly comparing all AI models remains a challenge due to the diversity of target proteins and categories within CASP competitions, as not all models compete across every category. Therefore, establishing a clear, comprehensive comparison is complicated.

### Model Evaluation Metrics

In CASP14, AlphaFold2 emerged as the top performer, achieving outstanding results across all targets in the test dataset. Specifically, it attained a median backbone accuracy of 0.96 ˚A RMSD, indicating that the deviation between predicted and actual atomic coordinates was limited to the typical length of a bond—demonstrating exceptional precision. Moreover, its all-atom accuracy reached an impressive 1.5 ˚A RMSD, and its GDT-TS score was 92.4, a clear testament to its structural prediction capabilities. RoseTTAFold, in second place, achieved a median backbone accuracy of 2.8 ˚A RMSD and all-atom accuracy of 3.5 ˚A RMSD, marking a solid, though comparatively lower, performance. Although AlphaFold2 did not participate in CASP15, many models, including RoseTTAFold and I-TASSER, leveraged insights gleaned from AlphaFold2 to refine their approaches. Given that CASP competitions are held every two years, evaluation results for AI models in protein structure prediction are not updated in real-time. To address this, the Continuous Automated Model Evaluation (CAMEO) contest provides weekly assessments of AI models, employing its own evaluation metrics, such as the LDDT. As of the end of June, models like OpenComplex and Swiss-Model led the CAMEO rankings, with LDDT scores of 81.7 and 79.2, respectively. These ongoing evaluations offer valuable insights into the continuous improvement of AI-based protein structure prediction models.

## End Users of AI Models

The development of AI-driven protein structure prediction models has revolutionized the field of computational biology, offering transformative opportunities across various domains. As highlighted in Section 2, these models have the potential to significantly enhance productivity and deepen our understanding in several key areas. Firstly, academic researchers in structural biology and related fields can leverage these AI models to advance their research by gaining deeper insights into protein functionality, uncovering new mechanisms, and exploring biological pathways. These models not only streamline the research process, reducing the need for costly and time-consuming laboratory experiments, but also pave the way for the development of next-generation models and methodologies. In collaboration with researchers, computational biologists and bioinformaticians can enhance their comprehension of complex proteomic landscapes and cellular functions.

Their work can lead to the development of innovative models that have broad applications, from commercial ventures to driving progress in drug discovery and drug design. A third group, encompassing pharmaceutical and biotechnology companies, as well as healthcare institutions, stands to benefit greatly from the insights derived from AI-driven protein structure predictions. These entities can harness research outcomes to develop novel therapeutic products, driving commercial success, while also advancing personalized medicine approaches tailored to individual patient needs. Finally, the ultimate beneficiaries of these advancements are patients and individuals affected by critical diseases or other biologically-driven conditions. Through the understanding of protein folding and protein-protein interactions, new therapies, drugs, and medical treatments can be developed, offering tangible improvements to their health and quality of life. The potential for AI models to impact disease treatment and management represents a profound shift in the future of medicine and healthcare.

## Limitations of Models

Despite the remarkable success of AF2 and other similar models in solving the protein-folding problem, several key limitations remain. While AF2 performs well in predicting a protein's single structure, it is unable to capture the dynamic nature of proteins, which often exist in multiple conformational states. These conformational changes are crucial for many biological functions, and the inability to predict them limits the model's capacity to fully characterize the diverse range of structural forms that proteins can adopt under different physiological conditions [34]. This oversight diminishes the model's utility in applications requiring an understanding of protein flexibility and functionality. Moreover, even with the advancements in AF3, which includes some improvements in predicting protein-protein interactions, the state-of-the-art models still struggle to accurately represent these complex and transient relationships. Protein-protein interactions are essential for cellular processes and drug design, yet current models have yet to reach the level of precision necessary for reliable predictions in these areas. Another significant limitation is that AlphaFold 2 and similar models are unable to predict post-translational modifications, which occur after a protein is folded [35]. PTMs, such as phosphorylation or glycosylation, play a crucial role in regulating protein activity, stability, and interactions, and their accurate identification is vital for understanding disease mechanisms and developing targeted treatments. Since AF2 relies solely on the amino acid sequence input, it cannot account for these important modifications, thereby limiting its application in more advanced biochemical studies.

From a technical perspective, several challenges persist in the protein folding prediction process. The generation of MSA maps and the search for homologous templates are computationally demanding tasks, re- quiring extensive processing power and time, even when utilizing parallel and distributed computing systems. Additionally, the deep learning and neural network architectures underpinning models like AF2 are inherently complex and difficult to interpret. These models are not governed by a single, easily understandable equation but are the product of intricate interactions between numerous functions and neurons, making it challenging to trace the rationale behind specific predictions. This lack of interpretability complicates the task of validating and refining predictions, as the "black box" nature of these models leaves little insight into how decisions are made. Furthermore, the training process itself is critical to the model's performance. Biases in the training data, whether due to overrepresentation of certain protein types or sequence similarities, can affect the model's accuracy, especially for proteins or scenarios that differ from the training dataset. This can lead to model bias, where certain predictions are overfitted or less reliable, limiting the generalizability and robustness of the model across diverse datasets and protein families.

## Future Research

The limitations outlined above, combined with the vast potential applications of solving the protein-folding problem, provide key directions for future research. One of the most critical areas for the next generation of models is to move beyond the prediction of single protein structures and focus on integrating broader functionalities. Future models should aim to predict not only the structures of individual proteins but also the interactions within large protein complexes, as well as with other biomolecules such as RNA, DNA, ligands, and ions-all of which are frequently encountered in structural databases like PDB. By expanding

their scope to include these more complex molecular interactions, AI models could contribute to a deeper understanding of cellular processes and enable the development of more advanced therapeutic strategies and commercial products that address many pressing biological challenges [36]. Additionally, a crucial step for advancing biological research is the innovation of computational methods. This includes developing more efficient algorithms, databases, and ontologies designed to streamline the retrieval of biological information

and the optimization of neural network architectures. Implementing these improvements would greatly enhance the speed, accuracy, and cost-efficiency of AI-driven models, enabling researchers to explore more intricate biological systems without the constraints of computational limitations. These advancements could significantly reduce the time, effort, and financial resources required to run large- scale protein structure predictions, thus accelerating progress across a range of scientific and medical domains. By focusing on these multifaceted challenges, future research will play a pivotal role in addressing the open questions in biomolecular sciences and in pushing the boundaries of what is possible with AI in biological research.

## Conclusion

The protein problem has garnered significant attention following groundbreaking innovations like AlphaFold, which have revolutionized the field of computational biology. However, it is still premature to declare the problem fully ”solved” in an absolute sense. While tremendous progress has been made, there remain numerous challenges to address, and continued advancements, refinements, and innovations are essential. Future developments hold great promise for uncovering new insights and unlocking a wealth of exciting discoveries with transformative applications that could fundamentally reshape our understanding of biology. These breakthroughs have the potential to not only enhance the field of computational biology but also significantly impact healthcare, drug discovery, and our broader approach to solving complex biological problems, ultimately improving our lives worldwide.

## References

1. Aebersold R, Agar JN, Amster IJ, Baker MS, Bertozzi CR, et al. (2018) How many human proteoforms are there? Nat Chem Biol 14(3): 206-214.
2. Ponomarenko EA, Poverennaya EV, Ilgisonis EV, Pyatnitskiy MA, Kopylov AT, et al. (2016) The size of the human proteome: The width and depth. Int J Anal Chem 2016: 7436849.
3. (2021) UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49(D1): D480-D489.
4. Chiang YT, Chien YC, Lin YH, Wu HH, Lee DF, et al. (2021) The function of the mutant p53-R175H in cancer. Cancers (Basel) 13(16): 4088.
5. Anfisen CB, Haber E, Sela M, White FH (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA 47(9): 1309-1314.
6. Finkelstein AV, Bogatyreva NS, Ivankov DN, Garbuzynskiy SO (2022) Protein folding problem: enigma, para- dox, solution. Biophys Rev 14(6): 1255-1272.
7. Ivankov DN, Finkelstein AV (2020) Solution of Levinthal’s paradox and a physical theory of protein folding times. Biomolecules 10(2): 250.
8. Sleator RD (2024) Solving the protein folding problem. FEBS Lett 598(23): 2831-2835.
9. Yang Z, Zeng X, Zhao Y, Chen R (2023) AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduct Target Ther 8(1): 115.
10. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873): 583-589.
11. Bramson J, Adler J, Dunger J, Evans R, Green T, et al. (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3 Nature 630(8016): 493-500.
12. Chen L, Li Q, Nasif KFA, Xie Y, Deng B, et al. (2024) AI-driven deep learning techniques in protein structure prediction. Int J Mol Sci 25(15): 8426.
13. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, et al. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373(6557): 871-876.
14. Pearce R, Li Y, Omenn GS, Zhang Y (2022) Fast and accurate ab initio protein structure prediction using deep learning potentials. PLoS Comput Biol 18(9): e1010539.
15. Wu R, Ding F, Wang R, Shen R, Zhang X, et al. (2022) High-resolution de novo structure prediction from primary sequence. bioRxiv.
16. Hie B, Candido S, Lin Z, Kabeli O, Rao R (2022) A high-level programming language for generative protein design. bioRxiv.
17. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M (2022) ProteinBERT: a universal deep- learning model of protein sequence and function. Bioinformatics 38(8): 2102-2110.
18. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31(13): 3381-3385.
19. Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins. Suppl 3: 171-176.
20. Zhou X, Zheng W, Li Y, Pearce R, Zhang C, et al. (2022) I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. Nat Protoc 17(10): 2326-2353.
21. Jimenez J, Doerr S, Martinez Rosell G, Rose AS, De Fabritiis G, et al. (2017) DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics 33(19): 3036-3042.
22. Zhang Y, Vass M, Shi D, Abualrous E, Chambers JM, et al. (2023) Benchmarking refined and unrefined AlphaFold2 structures for hit discovery. J Chem Inf Model 63(6): 1656-1667.
23. Edelmann MJ, Nicholson B, Kessler BM (2011) Pharmacological targets in the ubiquitin system offer new ways of treating cancer, neurodegenerative disorders and infectious diseases. Expert Rev Mol Med 13: e35.
24. Mok MT, Zhou J, Tang W, Zeng X, Oliver AW, et al. (2018) CCRK is a novel signalling hub exploitable in cancer immunotherapy. Pharmacol Ther 186: 138-151.
25. Weng Y, Pan C, Shen Z, Chen S, Xu L, et al. (2022) Identification of potential WSB1 inhibitors by AlphaFold modeling, virtual screening, and molecular dynamics simulation studies. Evid Based Complement Alternat Med 2022: 4629392.
26. Liang X, Zhang H, Wang Z, Zhang X, Dai Z, et al. (2022) JMJD8 is an M2 macrophage biomarker, and it associates with DNA damage repair to facilitate stemness maintenance, chemoresistance, and immunosuppression in pan-cancer. Front Immunol 13: 875786.
27. Pan X, Kortemme T (2021) Recent advances in de novo protein design: principles, methods, and applications. J Biol Chem 296: 100558.
28. Anishchenko I, Pellock SJ, Chidyausiku TM, Ramelot TA, Ovchinnikov S, et al. (2021) De novo protein design by deep network hallucination. Nature 600(7889): 547-552.

29. Rabbani G, Baig MH, Ahmad K, Choi I (2018) Protein-protein interactions and their role in various diseases and their prediction techniques. Curr Protein Pept Sci19(10): 948-957.

30. Evans R, ONeill M, Pritzel A, Antropova N, Senior A, et al. (2021) Protein complex prediction with AlphaFold-Multimer. bioRxiv.

31. Gomez Marin E, Posavec Marjanovic M, Zarzuela L, Basurto Cayuela L, Guerrero Martinez JA, et al. (2022) The high mobility group protein HMG20A cooperates with the histone reader PHF14 to modulate TGFβ and Hippo pathways. Nucleic Acids Res 50(17): 9838-9857.

32. Ivanov YD, Taldaev A, Lisitsa AV, Ponomarenko EA, Archakov AI (2022) Prediction of monomeric and dimeric structures of CYP102A1 using AlphaFold2 and AlphaFold Multimer and assessment of point mutation effect on the efficiency of intra- and interprotein electron transfer. Molecules 27(4): 1386.

33. McGuffin LJ, Aldowsari FMF, Alharbi SMA, Adiyaman R (2021) ModFOLD8: accurate global and local quality estimates for 3D protein models. Nucleic Acids Res 49(W1): W425-W430.

34. Wayment Steele HK, Ojoawo A, Otten R, Apitz JM, Pitsawong W, et al. (2024) Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 625(7996): 832-839.

35. Tikhonov D, Kulikova L, Rudnev V, Kopylov AT, Taldaev A, et al. (2021) A Changes in protein structural motifs upon post-translational modification in kidney cancer. Diagnostics (Basel) 11(10): 1836.

36. Lostao A, Lim K, Pallares MC, Ptak A, Marcuello C, et al. (2023) Recent advances in sensing the inter biomolecular interactions at the nanoscale - a comprehensive review of AFM-based force spectroscopy. Int J Biol Macromol 238: 124089.

**ISSN: 2574-1241**

DOI: 10.26717/BJSTR.2026.64.010069

**Theodoros Christodoulou.** Biomed J Sci & Tech Res



**Submission Link**: https://biomedres.us/submit-manuscript.php