International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Deep Learning Models for Protein Structure Prediction: A Comprehensive Survey

[1]C.S. Srushti, [2]P.M. Prathibhavani, [3]K.R. Venugopal

[1]Research Scholar, Department of Computer Science and Engineering, UVCE, Bengaluru, India.
[2]Assistant Professor, Department of Computer Science and Engineering, UVCE, Bengaluru, India.
[3]Retired Professor and Former Vice Chancellor, Bangalore University, Bengaluru, India.

[1]srushtichan@gmail.com, ORCID iD: 0000-0003-0450-1278,
[2]prathibhavani.pm@uvce.ac.in, ORCID iD: 0009-0009-8963-7158,
[3]venugopalkr@gmail.com, ORCID iD: 0009-0005-6091-8483.

**Corresponding author: C.S. Srushti*

Abstract: A major challenge in computational biology is determining the 3D structure of proteins from their 1D amino acid sequence. Accurate prediction of protein structures improves the understanding of protein function and facilitates applications in drug discovery, protein engineering, and structural biology. Modeling proteins using template-based approaches is difficult when homologous structures are unavailable in the Protein Data Bank (PDB). Recent advances in deep learning (DL) have significantly improved the accuracy of protein structure prediction (PSP). The current approaches employ convolutional neural networks (CNNs), recurrent neural networks (RNNs), graph neural networks (GNNs), and Transformer architectures to learn complex sequence patterns, evolutionary relationships, and long-range residue interactions that govern protein folding. By predicting long-range residue contacts, inter-residue distances, and geometric constraints, these approaches enable the generation of highly accurate three-dimensional protein structures. DL has transformed PSP, with models such as AlphaFold2, AlphaFold3, Boltz-1, and other recent architectures achieving near-experimental accuracy on many CASP targets. The Critical Assessment of Protein Structure Prediction (CASP) is the standard biennial community-wide benchmark for evaluating PSP methods. One of the biggest challenges remaining in this area is modeling regions of proteins that are intrinsically disordered and capturing the full range of protein dynamics and the relationships of different protein domains that exist in different proteins. This survey provides a comprehensive review of recent DL-based PSP methods, analyzes their architectural evolution, compares their strengths and limitations, and discusses emerging research challenges and future directions.

Keywords: Template-based modelling, Template-free modelling, Protein language models, AlphaFold, Protein structure prediction.

1 INTRODUCTION

Structural bioinformatics is a highly promising subject of study within computational biology. It involves the analysis and prediction of the three-dimensional structures of biological macromolecules, such as proteins, RNA (Ribonucleic acid), and DNA (Deoxyribonucleic acid) [1], [2]. Among biological macromolecules, proteins are prominent. Proteins are essential biological macromolecules involved in tissue development, maintenance, regulation, and repair. Proteins are essential for the normal functioning of all living organisms. As catalysts, proteins accelerate biochemical reactions involved in processes such as digestion, cell signalling, and DNA replication. Proteins also function as peptide hormones and antibodies and regulate several physiological processes, including nutrient transport, water balance, and muscle contraction. Proteins also contribute to the maintenance of acid–base balance.

Proteins are composed of twenty different amino acids arranged in a specific sequence. Under physiological conditions, amino acid (AA) sequences fold into distinctive three-dimensional structures. A protein's three-dimensional structure provides important insights into its biological function. It facilitates the understanding of catalytic activity, molecular transport, regulatory mechanisms, gene transcription, and molecular recognition [2]. Understanding the function of proteins is essential for developing effective diagnostic techniques [3], identifying and discovering drug targets [4], and therapeutic interventions [5]. The purpose of PSP is to predict the tertiary structure of a protein. The tertiary structure of individual polypeptide chains contributes to the formation of the quaternary structure of a protein. Only a small fraction of protein sequences has experimentally determined three-dimensional structures. The present paper reviews the current state of knowledge on PSP and highlights the limitations of existing methods.

1.1. Motivation

According to Anfinsen's thermodynamic theory, each protein's primary sequence, or chain of amino acids, has all the information necessary to determine how it folds.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

This amino acid sequence can be used to predict protein structure based on the minimal free-energy hypothesis [6]. Experimental methods include X-ray crystallography, Cryo-electron microscopy (cryo-EM), and Nuclear Magnetic Resonance (NMR) spectroscopy. Determining the structure of proteins requires expensive experimentation and investigations. Experimental determination of protein structures is expensive, labor-intensive, and time-consuming [7]. Each method has technical limitations that affect structure determination for certain proteins. Consequently, a considerable gap exists between the number of amino acid sequences identified through genomic projects and the number of proteins with experimentally determined three-dimensional structures.

Consequently, further research in computational PSP methods is required to reduce the gap between available protein sequences and experimentally determined three-dimensional structures. According to Levinthal's paradox, exhaustive exploration of all possible protein conformations is computationally infeasible because the conformational search space increases exponentially with protein length. This indicates that determining the best conformations by doing an exhaustive search would take a very long time, even for a short sequence. Several simplified formulations of the protein folding problem have been shown to be NP-complete or NP-hard in computational complexity theory [8], [9]. Consequently, the vast conformational search space remains computationally challenging to explore efficiently [10]. Towards this end, the objectives of the present research are:

i) Comprehensive review of DL models for PSP.
ii) Comparison of different techniques for PSP.
iii) Identifying research challenges in current PSP methods and future directions.

1.2. Deep Learning for PSP

Recent technological advances, such as the shift from classical statistical techniques to Machine Learning (ML) and DL techniques, have enhanced the power and accuracy of PSP. DL has become one of the most widely used technologies across a variety of scientific application domains. It enables the learning of representations of data at various degrees of abstraction via computer models made up of several processing layers. Through the backpropagation algorithm, it updates the internal model parameters used to compute progressively higher-level representations from preceding layers. DL techniques are also referred to as representation-learning methods because they learn multiple levels of data representation. Basic yet nonlinear modules progressively transform the input representation into increasingly abstract representations, and are combined to construct these models. By combining sufficient nonlinear transformations, highly complex functions can be learned. Due to advances in computing power and DL techniques, the field of PSP has advanced significantly [11]. Despite these remarkable advances, several important challenges remain before protein structure prediction can reliably address all classes of proteins and biological applications.

1.3. Research Challenges

Although DL-based PSP has achieved remarkable progress, several important research challenges remain. Current models primarily predict static protein structures, with limited capacity to model intrinsically disordered proteins, protein dynamics, membrane protein environments, and complex protein-protein interactions. Accurately predicting the effects of mutations and drug–protein interactions also remain challenging. Likewise, the black-box nature of transformer-based architectures limits biological interpretability and insight. The major research challenges are summarized below.

- **Intrinsically Disordered Proteins (IDPs):** Current prediction models perform well for proteins with stable tertiary structures but remain less effective for intrinsically disordered proteins, which exist as dynamic conformational ensembles rather than a single native structure.
- **Protein Dynamics**: Most prediction methods generate static structures and do not adequately capture conformational flexibility, folding pathways, or structural transitions occurring under physiological conditions.
- **Membrane Proteins:** Prediction accuracy remains comparatively lower for membrane proteins because of their complex lipid environments and the limited availability of experimentally resolved structures.
- **Protein–Protein Complexes:** Although recent multimer prediction methods have improved substantially, predicting transient interactions and large macromolecular assemblies remains challenging.
- **Mutation-aware Prediction:** Reliable prediction of structural changes caused by point mutations and disease-associated variants remains an active area of research.
- **Drug–Protein Interactions:** Current PSP models focus primarily on protein structures and generally do not directly model ligand binding, induced-fit effects, or protein–drug interactions.
- **Biological Interpretability:** Transformer-based architectures achieve excellent predictive performance but still provide limited biological interpretability, making it difficult to explain how specific structural predictions are derived.

Addressing these challenges is expected to improve the robustness, applicability, and biological relevance of future PSP methods.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

The review article is organized as follows: An introduction to PSP is provided in Section 1. The review methodology is presented in Section 2. Section 3 presents the survey statistics, while Section 4 describes protein structure and related concepts. Section 5 provides an overview of PSP, its importance, and the evaluation metrics used for comparison. Section 6 discusses DL-based ab initio prediction methods and Protein Language Models (PLMs). Section 7 presents a comparative analysis of different models. Section 8 discusses state-of-the-art PSP techniques beyond AlphaFold, including recent advances in PSP. The conclusion is presented in Section 9.

2 METHODOLOGY

To systematically assess recent advances in DL-based PSP, a structured review protocol was adopted, encompassing database querying, study selection, and thematic synthesis.

2.1. Search Strategy

A comprehensive literature search was conducted across major scientific databases, including PubMed, Scopus, and Web of Science, covering publications from 2015 to 2025. Keywords such as "*de novo* protein structure prediction," "*ab initio* protein modelling," "deep learning protein structure," and "neural networks in protein modelling" were used in various combinations. Boolean operators (AND/OR) were used to combine keywords during database searches. The reference lists of key articles and review papers were also screened to identify additional relevant studies.

2.2. Study Selection

Studies were included if they

i) Focused primarily on *de novo* or *ab initio* methods for PSP,
ii) Involved the use of machine learning or DL approaches, and
iii) provided either methodological innovations or benchmarking results.
iv) Review articles, research articles, and selected conference proceedings were considered.
v) Titles and abstracts were initially screened for relevance, followed by full-text review to confirm eligibility and exclude duplicate or irrelevant studies.
vi) Full texts were subsequently reviewed to confirm the presence of:
   a. A clear methodological description of the computational architecture (e.g., CNNs, GNNs, transformer models).
   b. Prediction outputs or evaluation metrics (e.g., contact precision, TM-score, torsion angles).
   c. Validation using benchmark resources such as CASP, CAMEO, CATH, or PDB-derived datasets.

2.3. Data Extraction and Thematic Synthesis

From the eligible studies, data pertaining to model architecture, input representation (e.g., raw sequence, multiple sequence alignment (MSA), co-evolutionary metrics), nature of the predicted output (e.g., binary contact maps, real-valued distance matrices, 3D coordinates), and quantitative performance metrics were extracted. Each model was mapped to one or more of the following thematic categories:

i) Contact-Based Prediction
ii) Inter-Residue Distance-Based Prediction
iii) Evolutionary Computation-Based Prediction
iv) Single-Sequence Prediction
v) Transformer-Based Architectures
vi) Diffusion-Based Models

These thematic groupings facilitated the synthesis of architectural trends, comparative evaluations, and insights into generalizability, interpretability, and biological relevance. Additionally, key innovations such as attention mechanisms, equivariant modelling, and geometric deep learning architectures were highlighted where applicable.

2.4. Exclusion Criteria

Studies were excluded if

i) They solely addressed algorithmic optimization without structural output.
ii) Studies published before 2015 were excluded unless they represented foundational contributions to the field.
iii) Papers that did not report standard structural evaluation metrics (e.g., RMSD, TM-score, GDT-TS, or lDDT) were excluded.
iv) Studies using identical benchmark datasets without novel methodological contributions were filtered out.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Fig. 1 illustrates the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram of the study selection process.

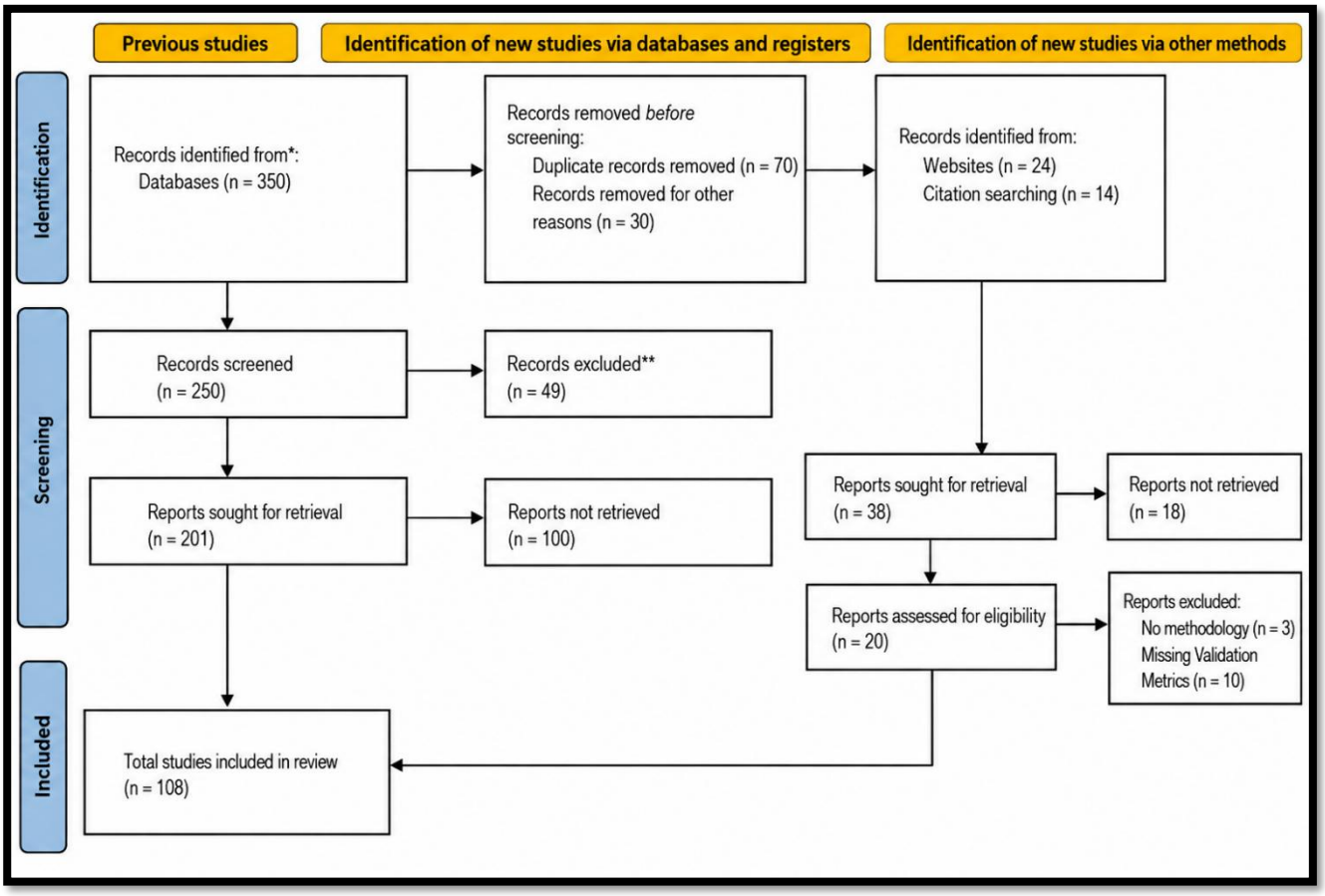


Fig. 1. PRISMA Flow Diagram of the study selection process

3 SURVEY STATISTICS

3.1. Taxonomy of DL-Based PSP Methods

PSP methods have evolved from contact-based approaches to advanced end-to-end DL frameworks capable of directly predicting three-dimensional protein structures. Based on the underlying prediction strategy and architectural characteristics, the reviewed methods are organized into distinct thematic categories. Fig. 2 illustrates the proposed taxonomy of DL-based PSP methods discussed in this survey. The taxonomy groups representative methods into contact-guided prediction, inter-residue distance prediction, end-to-end prediction pipelines, evolutionary computation, protein language model-based approaches, diffusion and next-generation models, and other emerging techniques.

3.2. Distribution of Surveyed Papers

The distribution of the surveyed papers across different publishers provides an overview of the primary publication sources contributing to recent advances in PSP. Fig. 3 depicts the distribution of the surveyed papers across different publishers. The x-axis represents the publishers, while the y-axis indicates the number of surveyed papers. Fig. 3 indicates that the surveyed literature is distributed across a wide range of publication sources, reflecting the multidisciplinary nature of PSP research. Traditional scientific publishers contribute a substantial proportion of the reviewed studies, while preprint repositories such as bioRxiv and arXiv also represent important sources for disseminating recent advances in this rapidly evolving field. This diverse publication landscape demonstrates the broad research interest in DL-based PSP across computational biology, bioinformatics, and artificial intelligence.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

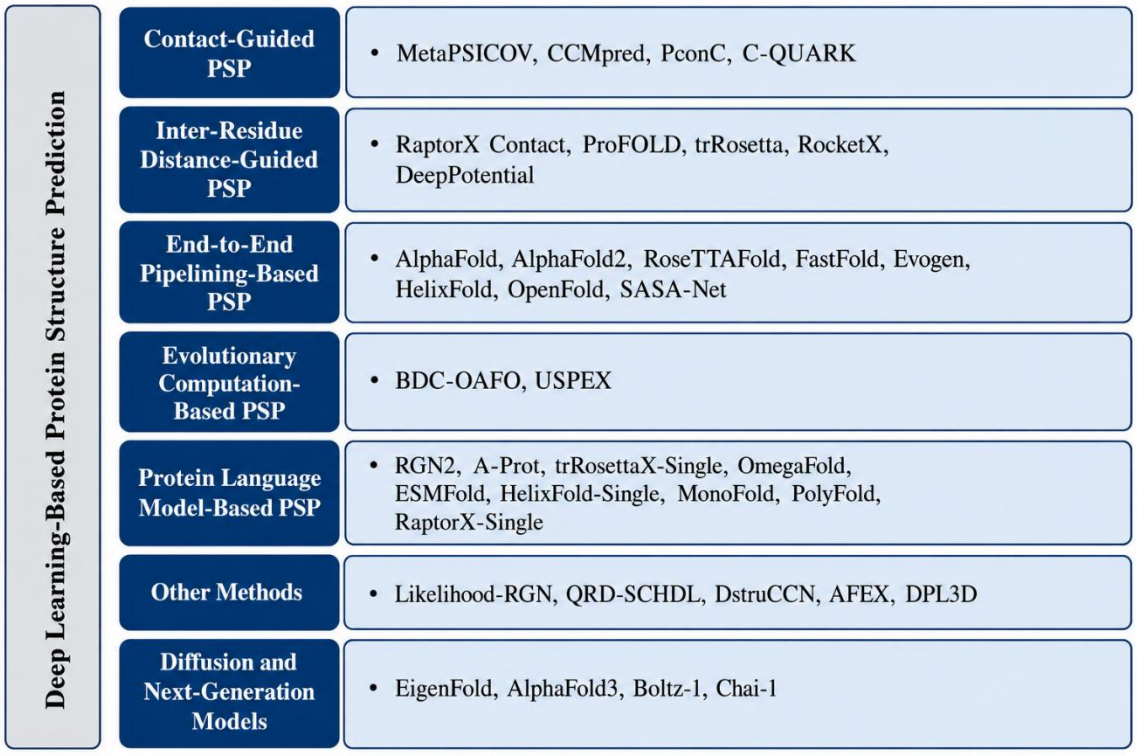


Fig. 2. Protein Structure Prediction Models

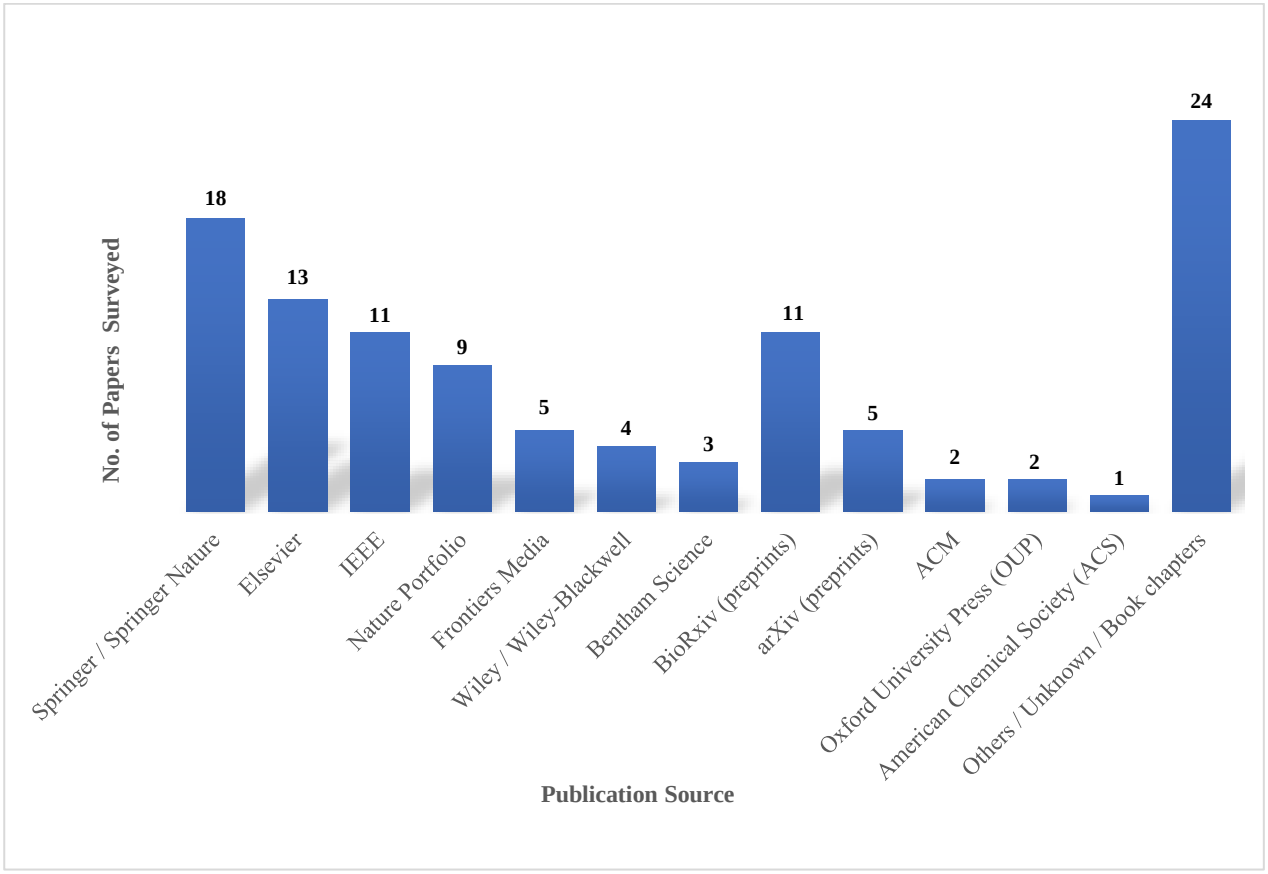


Fig. 3. Distribution of the Surveyed Papers Across Different Publication Sources

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

4 PROTEIN

Max Perutz and John Kendrew [12] determined the first three-dimensional (3D) structures of globular proteins. Structural bioinformatics examines the relationship between amino acid sequences, three-dimensional structures, and protein functions, thereby improving the understanding of genotype–phenotype relationships. Understanding the relationship between genotype and phenotype enables applications such as designing proteins that bind to specific targets, catalysing novel chemical reactions, and advancing biotechnology and medicine, including genome editing using CRISPR-Cas9.

4.1. Protein Representation

The historical development of protein and peptide chemistry—from the early identification of amino acids to advances in peptide synthesis—has been comprehensively reviewed in [13]. The 20 standard amino acids that constitute proteins share a common chemical structure. Their differing side groups cause them to behave differently. As a result, different combinations and sequences of amino acids give rise to proteins with diverse structures and functions. Every amino acid (AA) consists of a carboxyl group (COOH), an amino group (NH_2), a Cα atom, and an R group. The R group differs among amino acids, resulting in properties such as neutrality, acidity, basicity, aromaticity, or sulphur-containing side chains. The general structure of an amino acid is shown in Fig. 4. During peptide bond formation, the amino group of one amino acid reacts with the carboxyl group of another, releasing a water molecule and forming a peptide bond. The peptide bond formation is shown in equation (1).

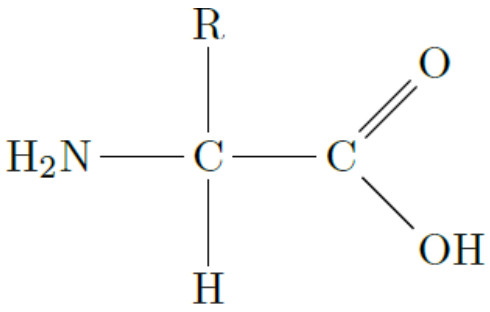


Fig. 4. General Structure of an Amino Acid

$$R_1\text{-}CH(NH_2)\text{-}COOH + R_2\text{-}CH(NH_2)\text{-}COOH \rightarrow R_1\text{-}CH(NH_2)\text{-}CO\text{-}NH\text{-}CH(R_2)\text{-}COOH + H_2O \quad (1)$$

During peptide bond formation, the carboxyl group of one amino acid reacts with the amino group of another amino acid, resulting in the formation of a covalent peptide bond with the release of a water molecule. This condensation reaction links individual amino acids into peptide chains, which subsequently fold into specific three-dimensional structures. The amino acid sequence connected through peptide bonds forms the primary structure of a protein and serves as the foundation for higher levels of protein organization.

4.2. Protein Structure

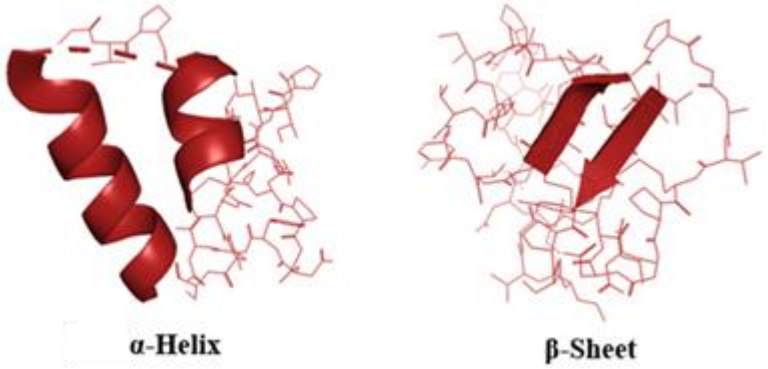
Proteins exhibit different levels of structural organization. Fig. 5(a)–5(d) illustrate the four levels of protein structure. Fig. 5(a)–5(c) illustrates the primary, secondary, and tertiary structures of the 1CRN protein. The primary structure of a protein consists of the linear sequence of amino acid (AA) residues, as shown in Fig. 5(a). Secondary structure is formed through hydrogen bonding between the backbone atoms of the polypeptide chain. Before the protein folds into the 3D structure also referred to as tertiary structure, secondary structural elements often form during the early stages of protein folding and contribute to the subsequent formation of the tertiary structure. Hydrogen bonds between backbone atoms give rise to two highly stable structural components within proteins. The two components are α-helices and β-sheets.

The 3D form of a protein is described by its tertiary structure. AA side chains can interact and link in a wide variety of ways. The interactions between side chains and bonding that take place within the protein itself determine its tertiary structure. Atomic coordinates describe the three-dimensional positions of atoms within the entire protein or a specific structural region. These chains can take on a wide variety of conformations since there are so many different ways to combine amino acids and rotate the chain at different points along the chain. Variations in the proteins' three-dimensional structure are caused by these conformational changes. Fig. 5(c) illustrates the tertiary structure of the 1CRN protein. In quaternary structure, two or more polypeptide chains assemble to form a functional protein complex, as illustrated in Fig. 5(d).

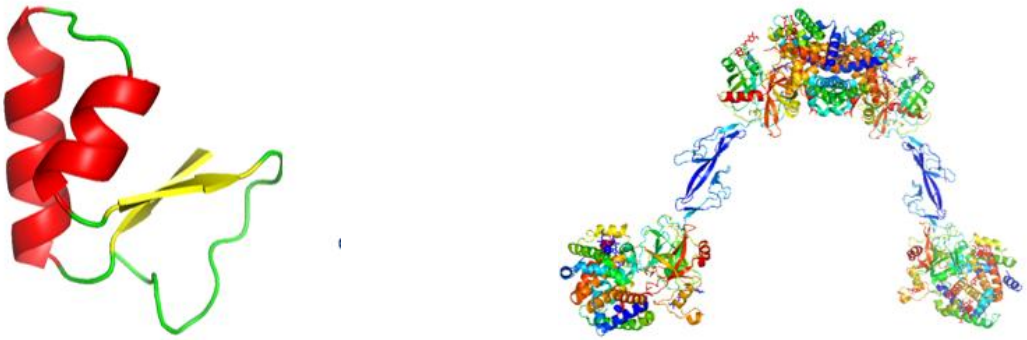
International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN

(a) Primary Structure of 1CRN



b) Secondary Structure of 1CRN



c) Tertiary Structure of 1CRN

d) Quaternary Structure of 4F49 Hb Protein

Fig. 5. Levels of Protein Structure

5 Protein Structure Prediction

Protein structure prediction can be categorized into one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) prediction tasks. One-dimensional and two-dimensional predictions are often used as intermediate steps toward accurate three-dimensional PSP. Multiple tertiary structures may further assemble to form quaternary structures [14]. According to Anfinsen's thermodynamic hypothesis [6], the native structure of a protein corresponds to its minimum free-energy state under physiological conditions, which has historically motivated many computational approaches to PSP. Protein free energy can be mathematically represented using functions that model interactions within the protein system, including non-bonded, hydrophobic, solvent, hydrogen-bonding, and entropic effects. The native protein structure corresponds to a specific arrangement of atomic coordinates. The native structure corresponds to the minimum value of the free energy function, which depends on the spatial arrangement of the atoms within the protein. PSP in the protein folding problem is seen as a global optimization problem since the energy functions are substantially nonconvex [14].

Anfinsen's hypothesis inspired the development of computational methods that evaluate protein conformations to identify the native state corresponding to the lowest free energy. The size of the conformational search space increases exponentially as a function of the protein sequences' length, which is the primary issue with this energy-driven technique. Proteins undergo both local and long-range folding interactions during the folding process. As proteins gradually adopt more complex conformations, simplified or coarse-grained models are often employed to guide the conformational search efficiently [15]. PSP can be broadly classified into experimental and computational approaches.

Experimental techniques include X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy, whereas computational methods employ algorithms, physics-based models, statistical potentials, machine learning, and DL to predict protein structures from sequence information and existing structural knowledge. Fig. 6 shows the overall growth in released protein structures per year. The x-axis represents the year, while the y-axis indicates the number of structures released annually [16].

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

The number of new protein structures that were made publicly available in a given year is represented by each bar on the graph. Noticeable increases in the number of released structures coincide with periods of significant advances in structural biology and computational prediction, including developments in cryo-electron microscopy and recent DL-based methods.

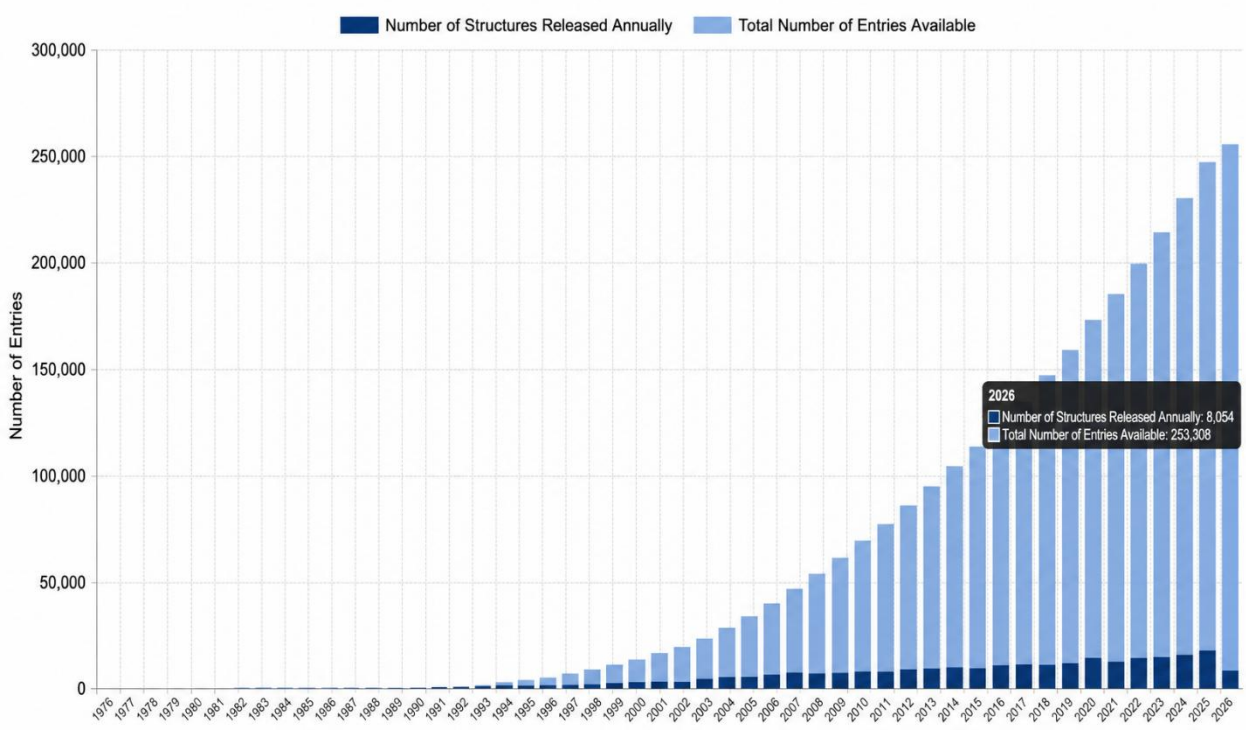


Fig. 6. Overall Growth of Released Structures Per Year

These challenges have motivated the development of diverse computational approaches for PSP, which are discussed in the following subsection.

5.1. Computational Prediction Methods

Although whole-genome sequencing has become inexpensive, experimental determination of protein structures remains costly. Consequently, computational prediction of protein structures has become essential. This need persists despite significant advances in experimental structure determination techniques, sequencing technologies, and the associated reduction in costs. These developments have been discussed in previous studies [17], [18]. Computational PSP methods can be broadly classified into the following categories:

i) *ab initio (de novo)* methods, in which protein structures are predicted from scratch using the first principles of physics.
ii) Template-driven approaches that use information from proteins with experimentally determined three-dimensional structures. This method can be additionally categorized into homology modelling, comparative modelling methods, and fold recognition methods. Homology modelling is based on sequence similarity between the target protein and the template, whereas fold recognition methods identify structural similarities despite low sequence identity.
iii) Hybrid approaches which uses both *ab initio* and template-based methods.

Template-Based Modelling (TBM)

TBM builds models by replicating and fine-tuning the structural blueprints of other, related proteins which are referred to as templates available in PDB. TBM is generally effective when the query and template share moderate to high sequence identity (typically above ~30%), although successful modelling is also possible below this threshold depending on structural conservation. Homology modelling primarily relies on sequence similarity, whereas threading and fold-recognition methods identify structurally compatible templates even when sequence similarity is low.

Template Free Modelling

Template-free modelling does not require experimentally determined structural templates from the PDB. This approach relies on conformational sampling and an effective ranking criterion. Conformational sampling is used to generate candidate models, while ranking criteria identify native-like structures. The three factors are necessary for effective *ab initio* modelling. The factors are:

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

i) An accurate potential energy function that discriminates native-like structures from decoy conformations. The accurate structure will have a state that is the most thermodynamically stable.
ii) A powerful search method that uses conformational search to efficiently find the low-energy states.
iii) Selection of native-like structures from a pool of decoy models [19].

Fig. 7 illustrates the general steps in TFM based PSP methods. By beginning with a target sequence, an MSA is generated by scouring a sequence database for sequences that are homologous. The sequence profile is then transformed from MSA. This sequence profile is later used to forecast the Secondary Structure Prediction (SSP), Solvent Accessibility (SA) and Backbone Torsion Angles (BTA) which are also known as local structural features.

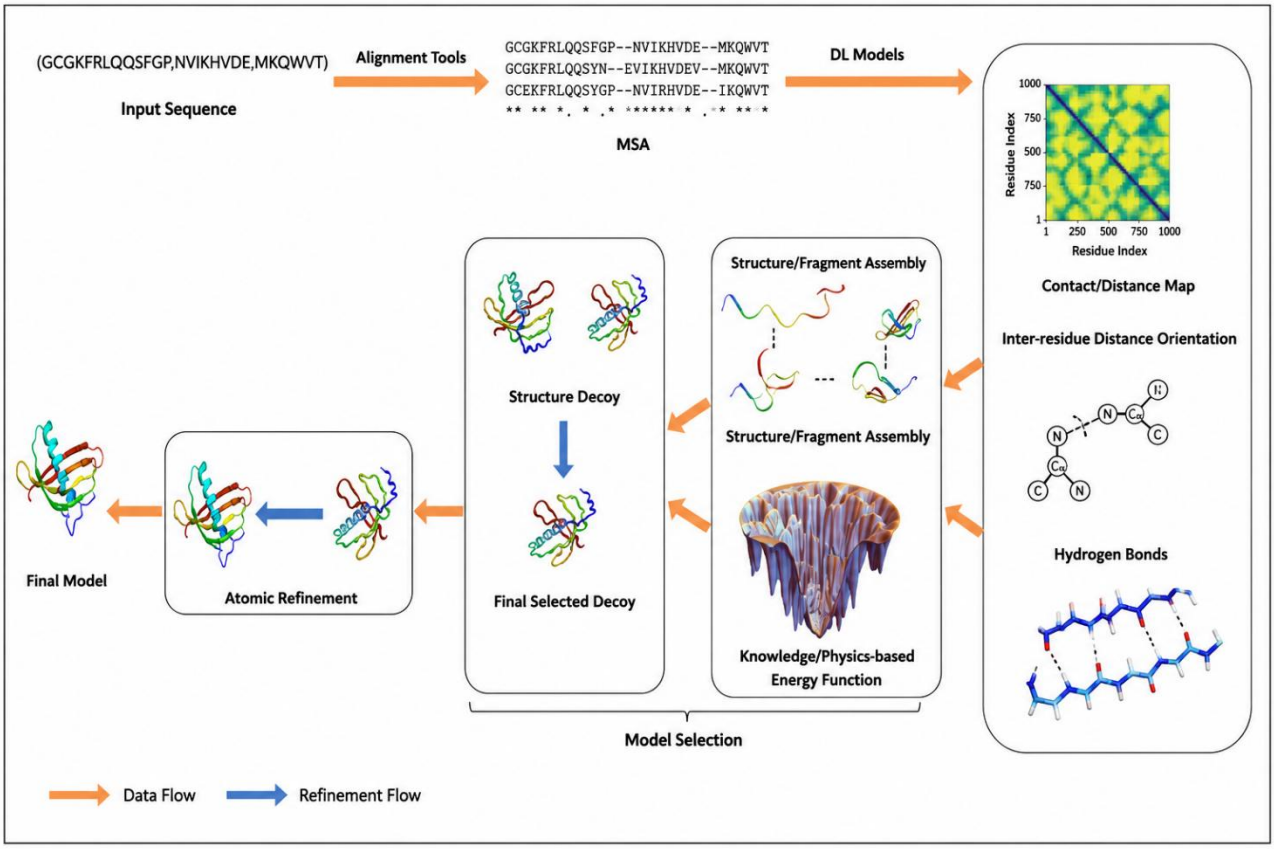


Fig. 7. General workflow of DL-based Template-Free PSP.

In modern PSP pipelines based on fragment assembly, a fragment library is searched for high-scoring using the structural features and the sequence profile. Threading uses TBM to detect structural templates. Co-evolutionary data from MSA feeds into a deep network to predict spatial features like hydrogen bonds and residue contacts. A hybrid energy function and DNN guide structure assembly. The lowest-energy models are refined at the atomic level, and representative structures are selected.

5.2. Review Articles Referred

A summary of the review articles related to PSP considered in this survey is presented in Table 1. The review articles summarized in Table 1 illustrate the rapid evolution of PSP research over recent years. Earlier reviews primarily focused on traditional computational approaches and protein structure annotations, whereas more recent studies emphasize DL, transformer-based architectures, and AlphaFold-inspired methodologies. This progression highlights the increasing role of artificial intelligence in advancing PSP and provides the foundation for the detailed analysis presented in the subsequent sections. Unlike the previous review articles summarized in Table 1, the present review provides an updated analysis of recent advances in DL-based PSP, including transformer architectures, protein language models, and diffusion-based models.

5.3. Measures for Evaluating Predicted Protein Structures

Several evaluation metrics have been developed over the past few decades to assess the structural similarity between predicted and reference protein structures. The following metrics are among the most widely used.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Table 1. List of Review Papers Referred

Authors	Description	Publisher	Citations
Kuhlman et al. [20]	Describes modern methods for predicting and designing proteins, and advantageous applications.	Elsevier	1093
Won et al. [21]	Provides an evaluation of accuracy estimation for CASP13 protein model structures	Wiley	69
Wen et al. [22]	Summarizes DL applications for analysing proteomics data.	Wiley	216
Torrisi et al. [15]	Evolution of predicting approaches for 1D and 2D Protein Structure Annotations, development of databases	Elsevier	377
Pearce et al. [23]	Reviews DL techniques for PSP and protein design which are two inverse processes in protein folding	Elsevier	182
Lain et al. [24]	Provide an overview of the novel DL approaches extensively employed in the last 2 years and widely used in CASP14	Wiley	48
Callaway [25]	Provides insight on the future of AI in protein-folding revolution	Springer	202
Huang et al. [26]	Insight on how research paradigms for PSP have shifted from data modelling to algorithmic modelling.	Elsevier	87
Bertoline et al. [27]	PSP methods before and after the introduction of AlphaFold	Frontiers	357
Elfsson [28]	Provides insight into the innovations inspired by the introduction of AlphaFold in CASP15	Elsevier	122

i) Root Mean Square Deviation (RMSD): RMSD measures the average distance between corresponding atoms of two optimally superposed protein structures. Atoms such as O, C, N, and Cα also known as backbone heavy atoms or all atoms of the superposed structures or sometimes only Cα atoms can have the RMSD determined. The RMSD values are measured in Å and the formula is given below [29].

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d_i^2} \tag{2}$$

where n is the number of comparable atom pairs and d_i represents the distance between the i[th] pair of atoms. Without factoring the orientations of the side chains, the Cα and backbone RMSD offer a comparison of similarity. The RMSD should be zero for protein structures that are 100% similar. As the degree of similarity diminishes, the RMSD of related proteins rises. Predicted structures with 3 Å RMSD or less from their native structure are regarded as highly accurate models in the context of PSP and can be incorporated for numerous kinds of protein structure-based investigations. Additionally, the structures within a 5Å RMSD range can shed light on the general topology/fold of the protein structures. For the first few CASP competitions, the primary criterion for evaluation, which takes different forms, was the RMSD of predicted structure concerning the superposed reference structures. The limitations of RMSD include sensitivity to outliers in poorly predicted regions, insensitivity to missing residues, and a strong dependence on the structural superposition between the reference and predicted models [30].

ii) Global Distance Test Total Score (GDT-TS): One of the standard accuracy metrics in CASP is the Global Distance Test Total Score, also known as GDT-TS, to overcome the disadvantages of RMSD. It measures the average percentage of corresponding Cα atom pairs within distance cut-offs of 1 Å, 2 Å, 4 Å and 8 Å [31].

iii) Template Modelling Score (TM-score): Another widely used evaluation metric is the Template Modelling Score (TM-score). It is intended to address two key issues with traditional measures like RMSD. Primarily, the TM-score assigns a higher weight to smaller distance errors than to bigger ones, which means that the score value is more sensitive to global fold similarity than to local structural changes. Second, the distance errors are normalised by the TM-score using a length-dependent scale. The TM-score is a numerical value between 0 and 1, where 1 denotes an exact match between two structures [32]. The formula to calculate the score is provided in [33] as

$$TM-score = \max\left[\frac{1}{L_{target}}\sum_{i}^{L_{common}} \frac{1}{1+\left(\frac{d_i}{d_o(L_{target})}\right)^2}\right] \tag{3}$$

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

where L_{target} is the number of residues in the reference (target) structure, L_{common} is the number of aligned residue pairs used in the comparison, d_i is the distance between the i^{th} pair of aligned residues after structural superposition, and $d_0(L_{target})$ is the length-dependent normalization factor.

$$d_0(L_{target}) = 1.24\sqrt[3]{L_{target} - 15} - 1.8 \quad (4)$$

Equation (4) defines the length-dependent normalization factor used in the TM-score calculation. This scaling function was derived by analysing large sets of related and unrelated protein structures and minimizes the dependence of the TM-score on protein length. This normalization substantially reduces the dependence of the TM-score on protein length. The major classes of structural comparison measures, including positional distance-based and contact-based metrics, are reviewed and analysed in [33]. The recent assessment techniques in the estimation of accuracy of protein models using DL are discussed in [21], [34].

iv) The Local Distance Difference Test (lDDT): The Local Distance Difference Test (lDDT), a score that does not consider superpositions (superposition-free score), assesses the local distance differences of all atoms in a model, including the validity of stereochemical plausibility. A single structure or an ensemble of equivalent structures might serve as the reference [35]. The lDDT score does not meet the mathematical requirements to be a metric, which constitutes a drawback. The same is true, though, for the majority of scores frequently used for structure comparison, such as GDT or RMSD based on iterative superposition when comparing models with a wide range of atoms. Among these evaluation metrics, RMSD, GDT-TS, TM-score, and lDDT are the most widely adopted benchmarks for assessing the accuracy of computational PSP methods, including those evaluated in CASP.

5.4. Critical Assessment of Protein Structure Prediction

The CASP is an international community-wide experiment for evaluating PSP methods and has been conducted biennially since 1994. Complete blind testing of structure prediction algorithms is the cornerstone of CASP. It provides researchers with an impartial assessment of state-of-the-art PSP methods and enables objective evaluation of computational modelling approaches. The basic objectives of CASP are

i) To evaluate the strengths and weaknesses of the current approaches to modelling protein structure from sequence.
ii) To identify areas where advancement is being achieved.
iii) To identify the major challenges limiting further progress in the field.

Table 2 summarizes the major CASP competitions, while Table 3 presents the prediction categories evaluated in each CASP edition.

Table 2. Summary of CASP Competition

CASP	Model QA	Model Quality Refinement
CASP1-CASP6	N	N
CASP7-CASP15	Y	Y

Table 3. Prediction Categories in CASP

CASP	Prediction							
	TS	SS	R-R Contact	Structure Complex	Disordered Region	Domain Boundary	Function	High-Accuracy Template-Based Modelling
CASP1	Y	Y	N	N	N	N	N	N
CASP2	Y	Y	N	Y	N	N	N	N
CASP3	Y	Y	N	N	N	N	N	N
CASP4	Y	Y	Y	N	N	N	N	N
CASP5	Y	Y	Y	N	Y	N	N	N
CASP6	Y	N	Y	N	Y	Y	Y	N
CASP7	Y	N	Y	N	Y	Y	Y	Y
CASP8	Y	N	Y	N	Y	N	Y	Y
CASP9	Y	N	Y	N	Y	N	Y	Y
CASP10	Y	N	Y	N	Y	N	Y	Y
CASP11	Y	N	Y	N	Y	N	Y	Y
CASP12	Y	N	Y	N	Y	N	Y	Y
CASP13	Y	N	Y	N	Y	N	Y	Y
CASP14	Y	N	Y	N	Y	N	Y	Y
CASP15	Y	N	Y	N	Y	N	Y	Y

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

The presence or absence of a category is represented by Y/N where yes is represented as Y and no is represented as N. Progress in protein structure modelling from 1994 to the present is illustrated in Fig. 8. The Y-axis (GDT-TS score) measures the accuracy of predicted protein structures compared to experimentally determined ones. A score of 100 means perfect alignment. The X-axis represents targets sorted by decreasing difficulty — easier ones on the left, harder ones on the right. Each coloured line represents the best-performing model from a specific CASP round (from CASP1 in 1994 to CASP15 in 2022). Prediction accuracy improved progressively from CASP1 to CASP13, followed by a substantial breakthrough in CASP14 with the introduction of AlphaFold2. Subsequent CASP editions have maintained near-experimental accuracy for many prediction targets.

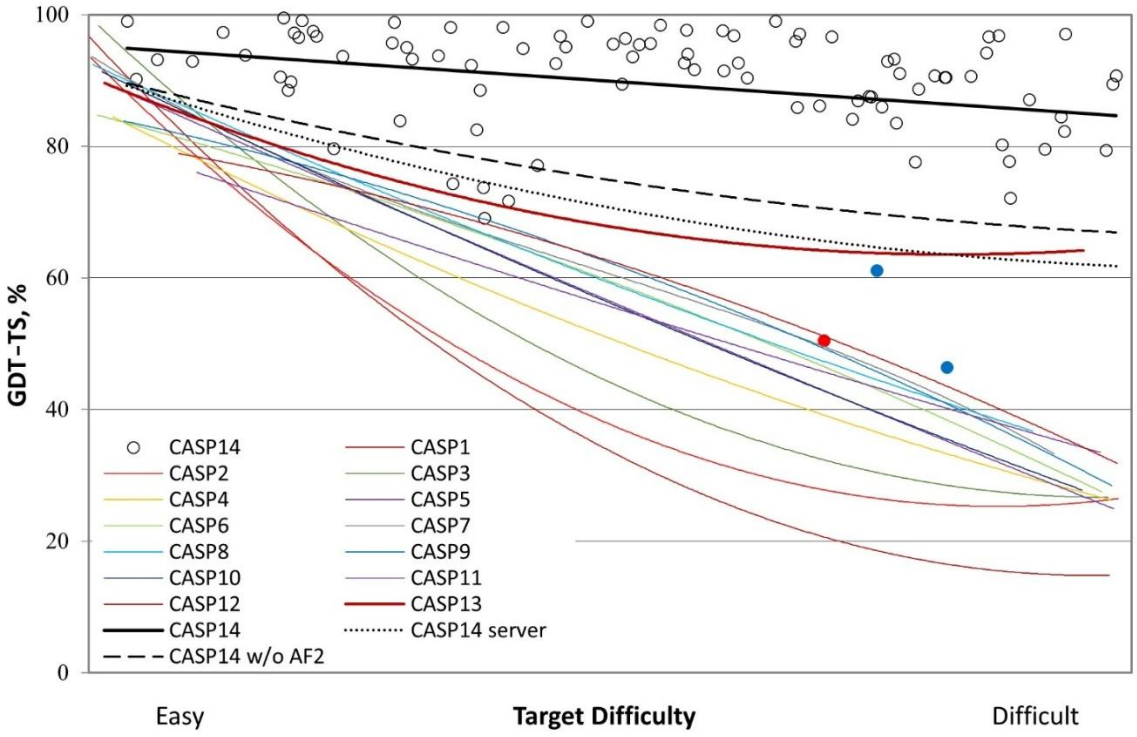


Fig. 8. Progress in Protein Structure Modelling

The progressive improvements observed across successive CASP competitions have largely been driven by advances in DL-based PSP methods, which are discussed in the following section.

6 AB INITIO METHODS FOR PSP

ab initio methods aim to predict a protein's 3D structure solely from its sequence without using templates. Over the past three decades, ab initio PSP has evolved from traditional physics-based simulations to modern DL frameworks capable of directly inferring three-dimensional structures from amino acid sequences. The major strategies employed in *ab initio* PSP are summarized in Table 4.

6.1. Deep Learning Architecture for PSP

The Feed Forward Neural Network (FFNN), also known as the Multilayer Perceptron (MLP), is one of the fundamental DL architectures. A single-layer perceptron performs linear classification by learning a weighted combination of the input features. Multiple layers enable neural networks to learn higher-order features. FFNNs have a feed-forward structure where information flows from input to output nodes, with each node connected to all nodes in the next layer [40]. In Long Short-Term Memory (LSTM) networks, gating is used to mitigate the vanishing gradient issue in RNNs, and as a result, long-term dependency can be learned. RNNs have demonstrated effectiveness in sequence-based applications such as protein structure prediction. In the Convolutional Neural Network (CNN), the architecture consists of convolution, pooling, and fully connected layers. The first layer in the CNN network is the convolutional layer in which learnable kernels are used to create feature maps from the input image [41]. Graph deep learning, especially Graph Convolutional Networks (GCNs), effectively models graph-structured data.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Table 4. Comparison of *ab initio* PSP Strategies

Method	Key Feature	Input	Accuracy Metrics	Strength	Computational Cost	Limitation
Physics-Based (AMBER [36], CHARMM [37])	Molecular dynamics using physical force fields	Primary sequence, force-field parameters	TM-score, GDT-TS, RMSD	Physically accurate	Increases with protein length and simulation time	Computationally expensive; force-field bias
Lattice-Based (Lau and Dill HP Model [38])	Folding on a discrete lattice	HP sequence, lattice type, contact energy	TM-score, GDT-TS, RMSD	Fast; suitable for theoretical studies	Reduced search space	Oversimplified for realistic proteins
Fragment Assembly [39]	Assembles structures from known fragments	Sequence, fragment library, MSA/PSSM	TM-score, GDT-TS, RMSD	High-resolution prediction	Thousands of decoys; Monte Carlo sampling	Dependent on fragment quality and scoring
Deep Learning	Learns sequence–structure relationships from large datasets	Sequence, MSA, optional structural templates	TM-score, GDT-TS, RMSD, lDDT/ pLDDT	Captures long-range interactions	Minutes to hours on GPUs	Limited for large complexes, ligand binding, and proteins with few homologs

These networks generalize convolution operations for non-Euclidean domains and are useful for solving complex problems like protein–protein interaction. A key hybrid model is the BRNN-CNN framework, where CNN kernels operate on BRNN memory segments. A notable example is the BLSTM-CNN, combining bidirectional LSTM and CNN elements. Optimizing such hybrids requires thoughtful architectural design, balancing accuracy, complexity, and computational demands. The frequently used DL architectures are illustrated in Fig. 9 [22]. Transformers use attention mechanisms to allow each input (e.g., a word or amino acid residue) to focus on all other components in a sequence, learning which ones matter most for a given task such as question answering or PSP. Unlike models that prioritize local interactions, transformers compute attention scores between all pairs of sequence elements and learn the relative importance of these interactions during training. This makes them particularly effective for protein folding, where both short- and long-range interactions are critical. Their success in natural language processing (NLP) has translated well into PSP, as illustrated in Fig. 10. A major challenge in applying neural networks to proteins lies in translating sequence data into usable real-valued features. One method for inter-residue contact prediction is to use a 2D matrix of residue-residue covariation scores derived from MSAs.

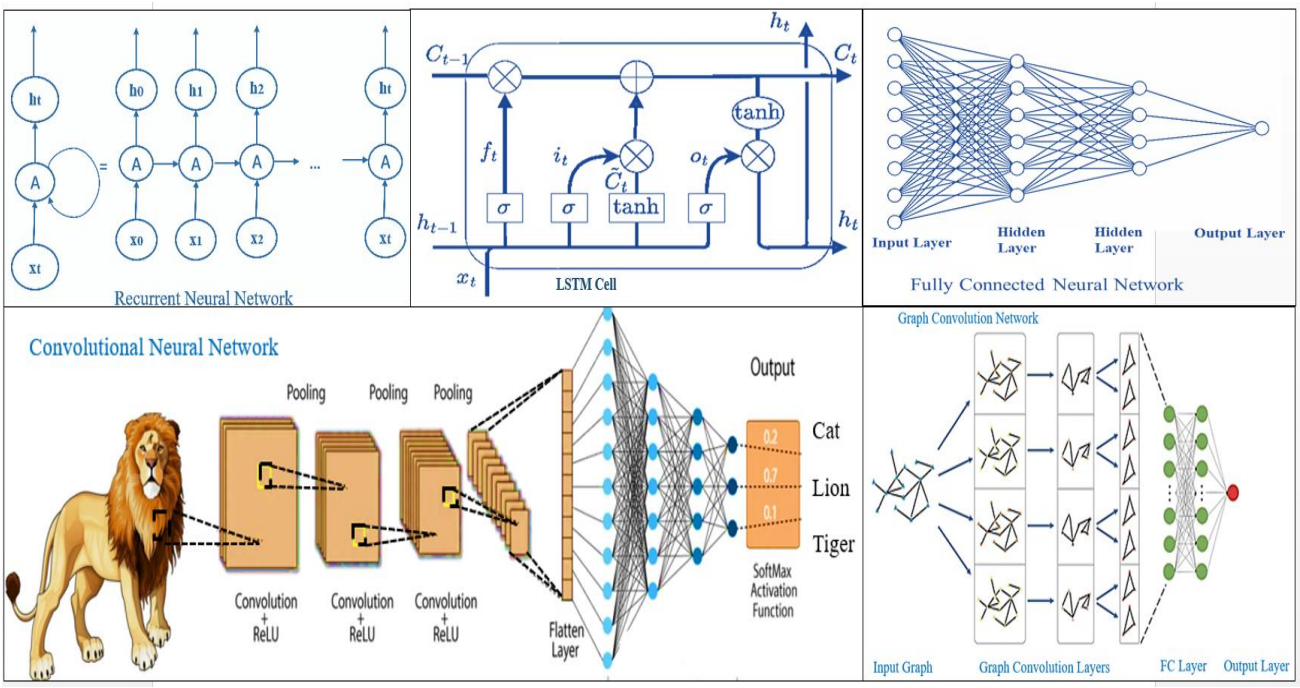


Fig. 9. Frequently used Deep Learning Architectures for PSP

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

The network learns to map these inputs to contact maps, which are also 2D matrices. By training on known examples, the network picks up recurring structural patterns to improve prediction accuracy.

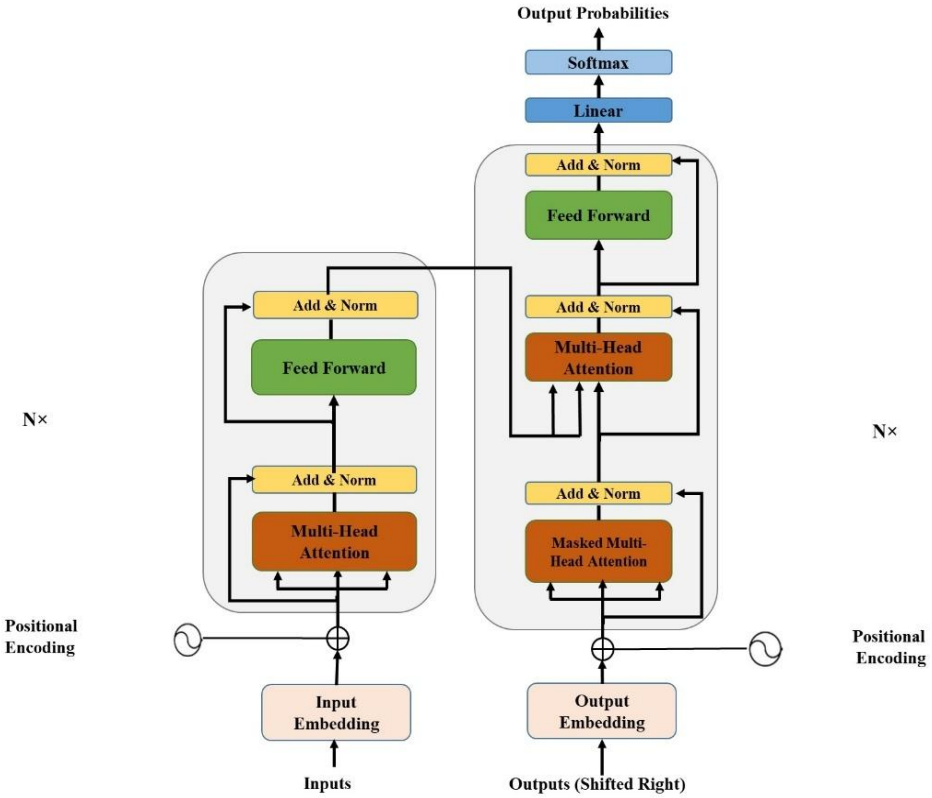


Fig. 10. Architecture of Transformer

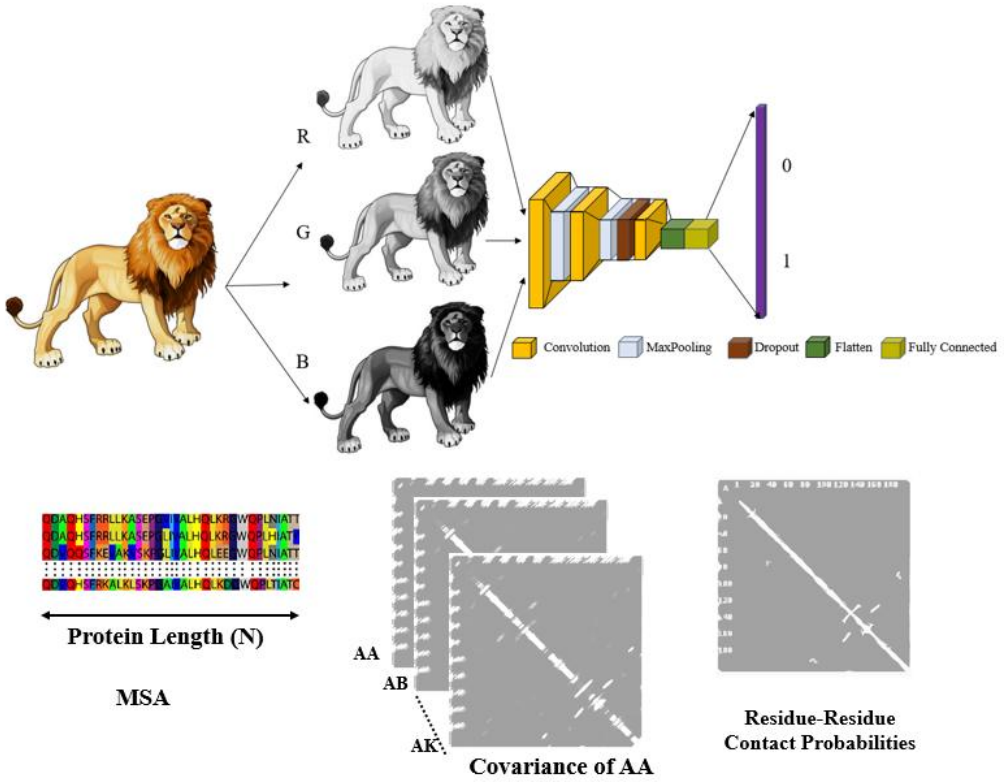


Fig. 11. CNN for Protein

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Fig. 11 illustrates the role of convolutional neural networks (CNNs) in processing protein features. The upper part of the figure uses an RGB image as a simple analogy to illustrate feature extraction through convolution, pooling, and fully connected layers. The lower part demonstrates how CNNs are applied to PSP by processing MSAs, covariance matrices, and residue-residue contact probability maps. Together, these inputs enable the network to learn structural patterns and predict spatial relationships between amino acid residues.

6.2. DL Contact Map-Guided PSP

A two-dimensional contact map of size $L \times L$, where L is the protein sequence length, contains a value of 1 when the distance between two residues is less than 8 Å and 0 otherwise. Estimating the contact probability for each pair of residues in a sequence is known as contact map prediction. Residue contacts are commonly categorized as local (<6 residues apart), short-range (6–11), medium-range (12–23), and long-range (≥24 residues apart in sequence). Contact prediction performance is commonly assessed by ranking the top L, $L/2$, or $L/5$ predicted residue pairs according to confidence scores and comparing them with the native structure. Local and short-distance contacts are usually not included in the evaluation. The input features used in the contact prediction problem have the shape $L \times L \times N$, where N is the number of pairwise descriptors.

These models provide accurate predictions with fewer decoys required for near-native conformations by integrating evolutionary coupling with DL. These advances enabled the development of *ab initio* structure prediction methods guided by predicted residue contacts. In 2012, a Deep Belief Network (DBN) was employed to forecast contact maps [42], and its performance was comparable to that of previous machine learning methods. Co-evolutionary analysis has gained increased attention as a result of the expansion of genomic data. For example, CCMpred [43] employs pseudo-likelihood maximization (PLM) for direct coupling analysis and is accelerated on GPUs and CPUs, making it suitable for large proteins and high-throughput analyses. Finally, by integrating three coevolution-based approaches with a conventional supervised machine learning approach, hybrid models such as MetaPSICOV [44] improve prediction by efficiently utilizing orthogonal sequence and structural information.

PconC [45] combines predictions from PSICOV and plmDCA [46] with alignments from HHblits and Jackhmmer at four different e-value thresholds, achieving approximately 20% improvement over individual methods. The combined approach yields contact sets distinct from individual predictors due to removal of low-confidence predictions or contacts detected by only one method, and the inclusion of intra-secondary structure interactions. C-QUARK [47] applies replica-exchange Monte Carlo simulations with multiple DL–based and coevolution-guided contact maps. It effectively handles targets lacking homologs using sparse, lower-accuracy contact predictions. Although contact maps significantly improved *ab initio* structure prediction, they provide only binary spatial information. This limitation motivated the development of inter-residue distance prediction methods, which offer richer geometric constraints and are discussed in the following subsection.

6.3. Inter Residue Distance-Guided PSP

Most inter-residue distance-based *ab initio* prediction methods follow a two-step strategy: first constructing a potential function from the predicted inter-residue distances, followed by three-dimensional structure generation through energy minimization. Although effective, these methods are limited by the accuracy of handcrafted potential functions and energy formulations. DL enables distance distribution prediction even with approximately 60 sequence homologs, allowing 3D model construction without extensive conformational sampling using geometric constraints [48]. ProFOLD converts predicted distances into a potential function to derive tertiary structure with minimal energy. Its core, CopulaNet, facilitates direct inter-residue distance computation from MSAs [49].

trRosetta, inspired by CASP13, is a web server leveraging DL and Rosetta. It predicts inter-residue geometries (distances and orientations) from AA sequences using a DNN, which are then used as restraints in Rosetta's energy minimization protocol [50]. RocketX introduces a feedback-based framework integrating GeomNet (geometry prediction), structure simulation, and EmaNet (model quality assessment) to improve PSP accuracy [51]. DeepPotential [52], another DNN-based model, uses AA sequences and co-evolutionary data from MSAs to predict structural descriptors for 3D modeling. Although DeepPotential demonstrated competitive performance, subsequent end-to-end models such as AlphaFold2 and RoseTTAFold achieved higher prediction accuracy by directly learning protein structures without relying solely on intermediate geometric representations. Table 5 provides the summary of PSP using Inter-residue distance.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Table 5. Comparison of Inter-Residue Distance-Based PSP Methods

Author	Method	Dataset	Contact Prediction Precision		CASP Performance	Advantages	Research Gap
			L/5, *L*/2 and *L* long-range residue contacts	RMSD/ TM-Score			
Xu et al. [48]	RaptorX-Contact (using distance geometry)	105 CASP11 test proteins, 76 past CAMEO hard targets, 398 membrane proteins	Not Reported	RMSD = 6.5 Å TM-Score 0.612	Top-performing contact predictor in CASP12 (F1 score on FM targets).	Best CASP12 F1 score on L/2 medium- and long-range contacts	Does not evaluate secondary structure or torsion-angle prediction
Ju et al. [49]	ProFOLD	Benchmark Dataset: CATH Test Dataset-CASP13	Precision of 0.840, 0.713 and 0.567	Avg TM-Score-0.662	Evaluated on CASP13 benchmark dataset.	Directly learns residue co-evolution from MSAs	Requires many homolog proteins
Du et al. [50]	trRosetta	CASP13	Not Reported	Avg TM-Score on CASP13 FM targets-~0.63	Achieved competitive performance on CASP13 FM targets.	Predicts inter-residue distance and orientation	High time complexity
Liu et al. [51]	RocketX	Training-DeepAccNet Testing- 483 benchmark proteins and CASP14 20 FM targets	Not Applicable	MAE-1.09 RMSE-1.50 Cβ-lDDT-0.677, RMSD-4.92 TM-Score-0.774	Evaluated on CASP14 FM targets; outperformed several template-free methods.	Feedback-guided geometry refinement	Increased runtime due to iterative feedback
Li et al. [52]	Deep Potential	Testing- 27 and 22 Free-Modeling domains from CASP13 and CASP14, 127 Hard targets from CAMEO	CASP13-0.854,0.751, 0.608 CASP14-0.902, 0.822, 0.687	Average TM-score of 0.672, 6.7% higher than trRosetta	Evaluated on CASP13 and CASP14 FM targets with competitive prediction accuracy.	Multitask prediction of distance, orientation, and hydrogen bonds	Relies on intermediate contact and distance representations

Inter-residue distance prediction represented a significant improvement over binary contact prediction by providing richer geometric constraints for protein folding. The methods summarized in Table 5 progressively improved structural accuracy through better distance estimation, orientation prediction, and DL architectures. These advances ultimately laid the foundation for end-to-end PSP frameworks, which are discussed in the following subsection.

6.4. End-to-End Prediction Pipeline for PSP

End-to-end PSP models directly infer three-dimensional protein structures from amino acid (AA) sequences by integrating multiple computational stages within a unified DL framework. Early end-to-end DL models include Recurrent Geometric Networks (RGN), which predicted backbone torsion angles for each residue without using co-evolution data, replacing traditional pipelines with a single differentiable approach [53]. NEMO [54] is another end-to-end model combining a coarse-grained neural energy function and unrolled simulation. NEMO offers faster inference, but drawbacks include slower learning, limited energy expressiveness due to Lipschitz regularization and gradient damping, and higher training/sampling costs compared to angle-predicting RNNs. AlphaFold [55], a prominent model using co-evolutionary data, employs thousands of CNN layers. It predicts inter-residue Cβ distances and torsion angles, constructs a statistical potential from distance distributions, and optimizes toward the native 3D structure using gradients. It performed best in CASP13. Despite its strengths, the original AlphaFold had limited performance for protein complexes, conformationally flexible proteins, and several challenging protein classes. AlphaFold2 (AF2) [56] demonstrated exceptional performance in CASP14, highlighting DL's potential in structure prediction.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Its strengths lie in (i) using transformers and attention mechanisms to capture long-range dependencies, (ii) applying symmetry principles to reason over 3D structures, and (iii) leveraging end-to-end differentiability for unified learning from protein data. AlphaFold2 represents proteins using MSA features and pairwise residue representations, enabling effective modelling of long-range spatial relationships. It uses two representations: MSA (evolutionary information) and pairwise residue representation. In the Trunk network, self-attention layers process MSAs, while pairwise features guide communication between residues. RoseTTAFold [57] adopts similar ideas, employing a 3-track neural network to integrate 1D sequence, 2D distance maps, and 3D coordinates, enabling accurate modelling of protein structures and protein-protein complexes through the joint processing of sequence, pairwise, and structural information. Given the high computational demands, ColabFold [58] offers a more accessible interface—via Jupyter or command line—by combining AF2/RoseTTAFold with MMseqs2 for accelerated prediction.

Other models build on AF2: FastFold [59], and Uni-Fold [60] improve training efficiency. Uni-Fold supports multimeric structure inference and achieves ~2.2× training speedup over AlphaFold2 (AF2) under similar conditions; EvoGen [61] enables MSA denoising or virtual MSA generation, supports AlphaFold2 (AF2) in low-data settings, enabling accurate folding even with single-sequence inputs and few-shot MSA. When combined with AF2, it facilitates probabilistic structure generation, capturing diverse sequence conformations. Additionally, its task-aware differentiable algorithm aids in applications like protein design; and re-implementations like HelixFold [62]and OpenFold [63] match AF2 accuracy. OpenFold provides a fully trainable open-source implementation while maintaining prediction accuracy comparable to AlphaFold2.

SASA-Net [64], a DL-based model, predicts protein 3D structures directly from inter-residue distances. Instead of atom coordinates, it represents structures via “residue poses”—coordinate systems fixed to each residue's backbone atoms. Its core architecture employs a spatially aware self-attention mechanism that adjusts poses based on all other residues’ features and predicted distances. End-to-end prediction pipelines have transformed PSP by eliminating many intermediate modelling stages and enabling direct learning from sequence information. The remarkable success of AlphaFold2 and subsequent architectures has established end-to-end DL as the dominant paradigm for accurate PSP, while ongoing research continues to improve computational efficiency, generalization, and applicability to complex biological systems.

6.5. Structure Prediction Based on Other Methods

The model proposed in [65] partitions the protein dataset into cluster subtrees and employs recurrent neural networks (RNNs) to capture local sequence-to-structure relationships. Once trained, the CRNN predicts structural properties such as distance matrices, backbone torsion angles, and secondary structure elements from the amino acid sequence. Recent advances use generative neural models for PSP. In [66], Variational Autoencoders (VAEs) show promising results for generating realistic 3D structures, provided sufficient high-quality training data is available. As generative models mature, their application scope will likely expand. DECO-VAE [67] is a variational autoencoder-based framework, uses data from structures generated by the Rosetta engine to learn latent protein structure distributions and generate tertiary conformations.

Attention-based DL models were introduced to protein sequence learning through transformer architectures [68], [69]. EquiFold [70] uses a coarse-grained, end-to-end differentiable representation that retains all-atom precision. It integrates geometric priors, including energy terms, to aid conformational landscape exploration. Although EquiFold demonstrates high speed and accuracy, its performance on completely novel protein folds has not yet been extensively evaluated. The model was primarily trained and assessed using challenging datasets such as antibody loops and mini-proteins. The growing success of attention mechanisms and generative learning has subsequently led to the emergence of protein language models, which learn structural and functional representations directly from large-scale protein sequence databases. These models are discussed in the following subsection.

6.6. Prediction of Protein Tertiary Structure using Protein Language Model

Recently, deep protein language models (PLMs) have transformed PSP and computational protein biology. Protein language models (PLMs) learn structural and evolutionary representations directly from large protein sequence databases, enabling structure prediction from amino acid sequences without requiring MSAs. PLMs encode amino acid sequences into distributed vector representations that capture structural and functional characteristics while enabling the assessment of evolutionary fitness across sequence variants. Discussion of current developments in protein language modelling and how they are applied to issues with protein property prediction down the line is presented in [71].

6.6.1. Single Sequence Structure prediction

It refers to a class of methods that predict a protein’s 3D structure using *only* its amino acid sequence—without relying on MSAs or evolutionary information. The list of PLM-based methods is summarized in Table 6. The modified evoformer module from AlphaFold2(AF2) is represented as M-Evfm in this study.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Table 6. List of PLM-Based Methods

Author	Method	PLM	Advantages	Limitations
Chowdhury et al. [72]	RGN2	AminoBERT	Outperforms AlphaFold2, RoseTTAFold, trRosetta on orphan proteins; up to 10^6× faster.	Overlooks complex, long-range interactions.
Hong et al. [73]	A-Prot	MSA Transformer	Requires fewer resources; trained using a single GPU; uses only sequence info.	Slightly lower TM-score than QUARK and Zhang-Server.
Wang et al. [74]	trRosettaX-Single	s-ESM-1b	Outperforms AF2 and RoseTTAFold on natural proteins; excels on design proteins.	Lower accuracy for single-sequence prediction of natural proteins.
Wu et al. [75]	OmegaFold	OmegaPLM	Requires only a single AA sequence (good for orphans/antibodies); 10× faster.	Cannot predict protein complex structures.
Lin et al. [76]	ESMFold	ESM-2	High-resolution accuracy; MSA not needed.	Slightly reduced accuracy in some segments compared to AlphaFold.
Fang et al. [77]	HelixFold-Single	Not specified	Fast prediction due to no MSA requirement.	Limited precision for proteins with few homologous sequences.
Barrett et al. [78]	MonoFold	ESM-2	Reduced computational burden; decent performance.	Lacks evolutionary info; less accurate.
Barrett et al. [78]	PolyFold	ESM-2 + MSA profile	Reduced training cost and improved performance.	Need to combines PLM embeddings with MSA profiles to improve prediction accuracy.
Jing et al. [79]	RaptorX-Single	ESM-1b, ESM-1V, ProtTrans	Outperforms AF2 on certain orphan PSPs; predicts single mutation effects.	Still struggles to predict correct conformation for many orphan proteins.

Table 7 summarizes representative review articles related to intermediate-stage PSP, including contact prediction, inter-residue distance prediction, and template-based approaches.

Table 7. List of Review Articles for PSP and Its Applications

Author(s)	Description	Prediction	Publisher	Citation
Adhikari et al. [80]	Explains the interactions between protein residues and how to predict these contacts.	Protein Contact Prediction	Springer	38
Skolnick et al. [81]	Examines how local vs. nonlocal physicochemical constraints influence native protein structures.	Physiochemical restraints	Elsevier	12
Zhang et al. [82]	DL-based PSP using templates.	Template-based PSP	BioMed Central	21

Beyond protein language models, several complementary computational approaches have also been explored for PSP. These methods are discussed in the following subsection.

6.7. Recent Advances in PSP

Several PSP methods have recently emerged, each incorporating distinct optimization strategies and DL architectures. Qi et al. [83] proposed Likelihood-RGN, which utilizes the ProteinNetX dataset and incorporates B-factors to improve global and local folding accuracy, though NMR model representation remains limited. QRD-SCHDL by V. Nallasamy et al. [84] integrates quantile regression with homolog-based DL for efficient predictions but still faces early-stage development challenges. Marchi et al. [85] introduced a multiobjective model using BRKGA and MUFOLD-CL for optimization, achieving competitive prediction performance without adaptive parameter control. Varanavasi et al. [86] presented Bartlett's Principal Regressive learning with Buffalo Optimization for robust PSP, although scalability improvements are needed. Shifana et al. [87] trained neural networks on AlphaFold data to predict atomic orientations efficiently but required high computational resources. Zhou et al. proposed DstruCCN [88], combining CNNs and supervised Transformers for single-sequence structure prediction, yet struggled with amino acid-nucleotide interaction binding. C.S. Srushti et al. [89] developed WebApp_Protein, integrating ESMFold and Streamlit for visualization and mutational scoring, though currently limited to single-chain protein analysis. Hu et al. introduced Cerebra [90], which accelerates training using parallel atomic coordinate prediction, but underperforms slightly compared to OpenFold.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Xie et al. [91] proposed AFEX, which customizes AlphaFold predictions with constraints to explore conformational diversity, though it lacks validation for large complexes. Chen et al. [92] proposed Protenix, which enhances AlphaFold3 with user-trainable modules and bug fixes but requires broader independent benchmarking. Lastly, Gao et al. introduced DPL3D [93], a platform to visualize mutant structures quickly, though it requires regular updates to stay effective. Although these approaches differ in their underlying architectures and optimization strategies, they collectively demonstrate the continuing evolution of computational techniques for PSP and motivate the exploration of evolutionary computation methods discussed in the next subsection.

6.8. Prediction Based on Evolutionary Computation (EC)

The optimization problem accompanying protein folding is high-dimensional and non-convex. EC techniques are well suited to this problem because they do not require gradient information, can escape local minima, and are readily adaptable to hybrid approaches such as force-field optimization and fragment assembly. Table 8 summarizes representative PSP methods based on evolutionary computation.

Table 8: List of PSP using EC

Author(s)	Method	Advantages	Research Gap
Silva et al. [94]	Self-adaptive Differential Evolution with fragment insertion	Improved results via parameter control and fragment insertion	Scalability issues with large protein sets
Boiani et al. [95]	GPU-accelerated 3D-AB off-lattice model	Enhances scalability and reduces execution time via GPU parallelism	Costly local search (Hooke–Jeeves Direct Search); limited by GPU shared memory
Perpinelli et al. [96]	Self-adaptive algorithm using Monte Carlo fragment insertion + conformational clustering	Competitive energy and RMSD results	Errors in predicted secondary structure reduce overall accuracy
Nallasamy et al. [97]	BDC-OAFO: Bingham CNN + Oppositional Artificial Fish Optimization	Tackles sensitivity issues with imbalanced data	Prediction accuracy remains inconsistent across different protein classes
Rachitskii et al. [98]	USPEX for global optimization in PSP	Finds deep energy minima effectively	Limited evaluation excluded proteins with cis-proline residues
Parpinelli et al. [99]	Dynamic speciation + Rosetta Quota fragment library + structure-guided selection	Rich fragment diversity and robust conformational exploration	Not tested on very large or CASP targets

Evolutionary computation remains an effective optimization framework for PSP, particularly when integrated with fragment assembly and physics-based energy functions. Recent research has increasingly combined evolutionary optimization with DL to improve conformational exploration while reducing computational complexity.

6.9. Diffusion-Based PSP Methods and Biomolecular Modelling

EigenFold [100] is the first diffusion-based generative modelling framework for protein structure and structural ensemble prediction, employing a “harmonic diffusion” process in which a protein sequence is represented as a set of harmonic oscillators. This approach injects structures into the eigenmodes of the system, resulting in a cascade-resolution generative process where the coarse global structure is sampled first, and then locally refined by a SE(3)-equivariant score model that predicts update “forces” The model is an MSA-free (single-sequence) model that takes primary protein sequences and pretrained node and edge embeddings from the OmegaFold Geoformer stack as input. In terms of accuracy, EigenFold achieves a median TM-score of 0.84, median GDT-TS of 0.79, and median Cα RMSD of 3.50 Å on CAMEO targets (less than 750 residues), on par with RoseTTAFold but still behind the accuracy of AlphaFold2 and ESMFold.

Inference is relatively fast for a diffusion model, requiring only 100 to 300 steps and the framework provides a built-in ranking based on an approximate Evidence Lower Bound (ELBO) to choose the most likely sampled conformations .The inability to predict conformational heterogeneity (e.g. fold-switching or ligand-induced conformational change) to high accuracy, a bias towards the single-structure output of the frozen OmegaFold embeddings, and a computational limitation that limits prediction to proteins under 750 residues are some of its research gaps. AlphaFold3 [101] drops the rigid-body frames and torsion-angle calculations of AlphaFold2’s “Structure Module”. Instead, it applies a generic All-Atom Diffusion Module. It is a universal representation of all physical entities, amino acids, nucleotides, ions and chemical modifications as point clouds.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

The network starts from the entire system as unstructured Gaussian noise in 3D space and through several steps performs denoising to predict accurate raw atomic coordinates. It simultaneously predicts complex, joint biomolecular interactions (Protein-DNA, Protein-RNA, Protein-Ligand and antibody-antigen interfaces) without requiring specialised separate tools. It avoids the need for explicit torsion-angle constraints by directly predicting atomic coordinates. But the iterative nature of reverse diffusion results in significantly slower inference than old deterministic models and therefore limits high-throughput virtual screening. The diffusion process may occasionally generate stereochemical inconsistencies, such as chirality violations, atomic clashes, or unrealistic molecular configurations, particularly when experimental constraints are limited. AlphaFold3 accepts diverse biomolecular inputs, including proteins, nucleic acids, small molecules, covalent bonds, and residue modifications, while also utilizing multiple sequence alignments (MSAs). On the CASP15 RNA targets, it achieved a full-complex GDT score of 86.9 for a ribosomal subunit and 90.1 for a specific transcriptional regulator.

The computational requirements include a high computational cost for challenging targets like antibody-antigen interfaces as they are trained in three stages with increasing crop sizes (384, 640, and 768 tokens) using a mini-batch of 256 samples. The model can occasionally produce chirality violations (4.4% in benchmarks) and atomic clashes, including rare instances where entire homomer chains are predicted to overlap. As a generative diffusion model, it may predict structured conformations in intrinsically disordered regions where experimental evidence remains limited. Boltz-1 [102] is an open-source DL model for predicting 3D biomolecular complex structures with AlphaFold3-level accuracy. The Boltz-1 workflow is composed of three stages: a) Data Processing and MSA Pairing: Processes amino acid sequences, SMILES(Simplified Molecular Input Line Entry System) strings and genomic sequences with a novel algorithm for protein interactions that takes into account sequence redundancy b) Core Architecture, with a trunk with Atom Attention Encoder and MSA Module, optimising attention bias for denoising c) Innovations that improve backpropagation and make efficient use of transformations, such as architectural reordering and residual transformer layers. Further the boundary of capabilities of these models with Boltz-steering, a new inference time steering technique that is able to fix hallucinations and non-physical predictions from the models.

The inputs to the model include amino acid sequences, SMILES strings and covalent bond information, Nucleic Acids. The results for CASP15 dataset with Mean LDDT of ~0.46, Mean TM-Score of ~0.49 whereas for Curated PDB Dataset, mean LDDT score was ~0.81 and Mean TM-Score was ~0.91. The model was optimized for speed through attention bias sharing and caching, which allows the trunk of the model to run once while sharing representations across multiple diffusion timesteps. The speed of computation was approximately one-quarter of the computation time required by AlphaFold3. The benchmarking and evaluation were performed on NVIDIA A100 80GB GPUs. The study highlights that the overlapping ligands were not removed during data processing, leading to misleading learning signals from the Protein Data Bank (PDB). This resulted in the model incorrectly interpreting overlapping ligands as normal configurations. Additionally, the model's training on token sizes of 384 and 512 was inadequate for complex biological structures, hindering its global perspective and causing failures in maintaining structural integrity in larger assemblies. Chai-1[103] is a GPU-accelerated, multi-modal biomolecular structure prediction model that infers 3D structures of proteins, nucleic acids, ligands, and complexes from sequence and SMILES inputs. The API accepts up to 5 molecules per request (proteins ≤1024 residues; DNA/RNA ≤3072 bases; ligands ≤128 characters) with optional per-protein MSAs from UniRef90, MGnify, or Small BFD. It returns CIF-format coordinates plus optional pLDDT and PAE scores, enabling single-sequence and MSA-informed workflows for drug discovery and protein/nucleic acid engineering.

The input to the model is Biopolymers, Small Molecules with MSA being optional For Protein Monomer Prediction, CASP15 for 69 Targets it achieves average lDDT = 0.849 with MSA. Latency is reduced due to distributed MSA search pipeline that runs in parallel across multiple shards of genetic databases. Model training reportedly required 128 NVIDIA A100 GPUs for approximately 30 days using a batch size of 128. The model occasionally gets the individual chains right but the wrong relative spatial orientations within a complex. It is very sensitive to modified residues. If these are replaced by standard analogues, the predicted structure can change significantly, because the model depends on these modifications for accurate modelling. Prediction of antibody-antigen interfaces at high accuracy is a challenging problem and even with restraints the fraction of “high quality” predictions is low (4-8%). Diffusion-based models represent the latest generation of PSP frameworks, extending prediction from individual proteins to complete biomolecular systems involving nucleic acids, ligands, and other molecular partners. Compared with earlier end-to-end architectures, these models provide greater flexibility in modelling complex molecular interactions while introducing additional computational requirements. The comparative performance of the major PSP models is discussed in the following section.

7 COMPARISON BETWEEN PSP MODELS

Table 9 summarizes the architectural characteristics of representative PSP models. Although these models share the common objective of predicting accurate three-dimensional protein structures, they differ substantially in their feature representations, refinement strategies, and network architectures. Fig. 14 further illustrates these architectural differences by comparing the major design components adopted in recent PSP frameworks.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Table 9. Comparison of Representative PSP Models

Method	Input	MSA Required	Primary Metric	Speed	Dataset	Inference Time	Computational Requirements
AlphaFold2	MSA	Yes	TM-score: 0.88 (CAMEO), 0.85 (CASP14); median **lDDT** ≈0.85	>10 min/protein	CASP14, CAMEO	~85 s (384 residues)	128 TPUv3 cores for training; high-end GPU/TPU for inference
RGN2	Single Sequences	No	1.85 Å better than AF2 on orphan proteins	~10^6× faster	ProteinNet12, ASTRAL	Very fast	AminoBERT trained on 512 TPU cores
A-Prot	MSA	Yes	TM-score: 0.658 (CASP13), 0.576 (CASP14)	Faster than AF2	CASP13, CASP14	Fast inference	NVIDIA RTX 8000 (48 GB), ~5-day training
trRosettaX-Single	Primary sequence	No	TM-score: 0.79 (orphan proteins)	~2× faster than AF2	CASP14	No MSA search	<10% of AF2 computational cost
OmegaFold	Single-Sequence	No	lDDT: 0.82; TM-score: 0.79	~10× faster than AF2	CAMEO, CASP	7.6–128 s	NVIDIA A100 GPU
ESMFold	Amino Acid Sequence	No	TM-score: 0.83 (CAMEO), 0.68 (CASP14)	6–60× faster than AF2	CASP14, CAMEO	14.2 s (384 residues)	Multi-GPU FSDP training
HelixFold-Single	Primary sequence	No	Median TM-score: 0.78–0.82	Very fast	CASP14, CAMEO	Very fast	Single NVIDIA A100 GPU
RaptorX-Single	Primary sequence	No	TM-score: 0.43; GDT-TS: 43.4	Fast	CASP14, Orphan proteins	No homolog search	ProtTrans (3B parameters), 32 GPUs

Fig. 12 compares the architectural characteristics of representative PSP models. Although these models share the common objective of predicting accurate three-dimensional protein structures, they differ considerably in their network architectures, feature representations, and refinement strategies. Recent models increasingly employ transformer-based architectures, attention mechanisms, and iterative refinement to improve prediction accuracy and computational efficiency. Table 10 compares the scope of the present survey with representative review articles published between 2019 and 2024. Earlier surveys primarily focused on specific aspects of protein structure prediction, such as contact prediction, AlphaFold2, or protein language models, whereas the present survey integrates contact prediction, inter-residue distance prediction, end-to-end architectures, protein language models, diffusion-based models, evolutionary computation, and comparative analyses, thereby providing a broader overview of recent advances in DL-based PSP.

Table 10. Comparison with Prior Surveys from 2019-2024

Survey	Contact Maps	AlphaFold2	PLMs	AlphaFold3	EC Methods
Kuhlman et al. [20]	Introduced	Not covered	Not covered	Not covered	Basic GA/PSO
Torrisi et al. [15]	Expanded	Not covered	Not covered	Not covered	Hybrid EC
Pearce et al. [23]; Laine et al. [24]	Detailed	Early coverage	Not covered	Not covered	Limited
Callaway [25]	Comprehensive	Benchmarked	Initial PLM mention	Not covered	Moderate
Huang et al. [26]; Bertoline et al. [27]; Elofsson [28]	Comprehensive	Widely covered	PLMs emphasized	Not covered	Comparative
Current Survey	Integrated taxonomy	Comparative benchmarks	Single-sequence PLMs	Covered	Hybrid EC-DL integration

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Feature	Architecture	Key Components	Refinement Strategy
AlphaFold2	Evoformer, Structure Module, Recycling	Axial attention, Invariant Point Attention	Iterative refinement
HelixFold Single	Transformer, DisentangledAttention Transformer, Geometric Modelling, Structure Module, Recycling	Residue interactions, Relative positions, EvoformerS, Invariant Point Attention	Iterative recycling
OmegaFold	OmegaPLM, Geoformer, GAU, RoPE, IPA	Structural and physical pairwise, Gated Attention Units, RoPE, IPA	Internal geometric updates, recycling-like iterations
RaptorX-Single	Sequence Embedding, Modified Evoformer, IPA	Scalar, point, and pair attention	Multiple network blocks iteratively update geometry
RGN2	AminoBERT, Geometric Module	Latent structural information, Cα backbone geometry	No deep recycling, refinement delegated to Rosetta
A-Prot	MSA Transformer, trRosetta	Evolutionary information, 2D distance and dihedral angle predictions	Refinement occurs during restraint optimization
trRosetta-Single	s-ESM-1b, Res2Net, Knowledge Distillation, Rosetta	2D geometry predictions, 3D folding via Rosetta energy minimization	Decoupled two-step process
ESMFold	ESM-2, Folding Blocks, Equivariant Transformer, Recycling	Evolutionary patterns, 1D language model, Equivariant transformer	Recycling in three steps

Fig. 12. Comparison of the Architectural Features of Representative PSP Models

8 Application of AlphaFold Predicted Tertiary Structures

Some of the common applications of protein structures predicted by AlphaFold are listed in Table 11. Computationally predicted protein structures generated by AlphaFold are made available through the AlphaFold Protein Structure Database (AlphaFold DB), providing researchers with predicted three-dimensional structures for a large number of proteins. Fig. 13 illustrates the AlphaFold-predicted three-dimensional structure of the protein Evasin P1126 visualized using the AlphaFold DB molecular structure viewer. The interface provides access to structural representations, chain information, quality assessment, and model export options available for the predicted protein model. The AlphaFold DB identifier for this protein is AF_AFA0A023FF81F1.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Table 11. Applications of AlphaFold Predicted Tertiary Structures

Domain	Application	Impact
Drug Discovery [104], [105], [106]	Predicts protein targets and binding sites for small molecules	Accelerates lead identification and rational drug design
Protein Engineering [107]	Guides design of novel proteins with desired functions	Enables synthetic biology and enzyme optimization
Protein-Protein Interaction [108]	Prediction of interaction interfaces and functional annotation	Improves understanding of cellular pathways, enzyme regulation, structural assembly, and disease mechanisms

Table 11 summarizes some of the major applications of AlphaFold-predicted protein structures across different domains of structural biology. The widespread availability of high-quality predicted structures has accelerated research in drug discovery, protein engineering, functional annotation, and the investigation of biomolecular interactions.

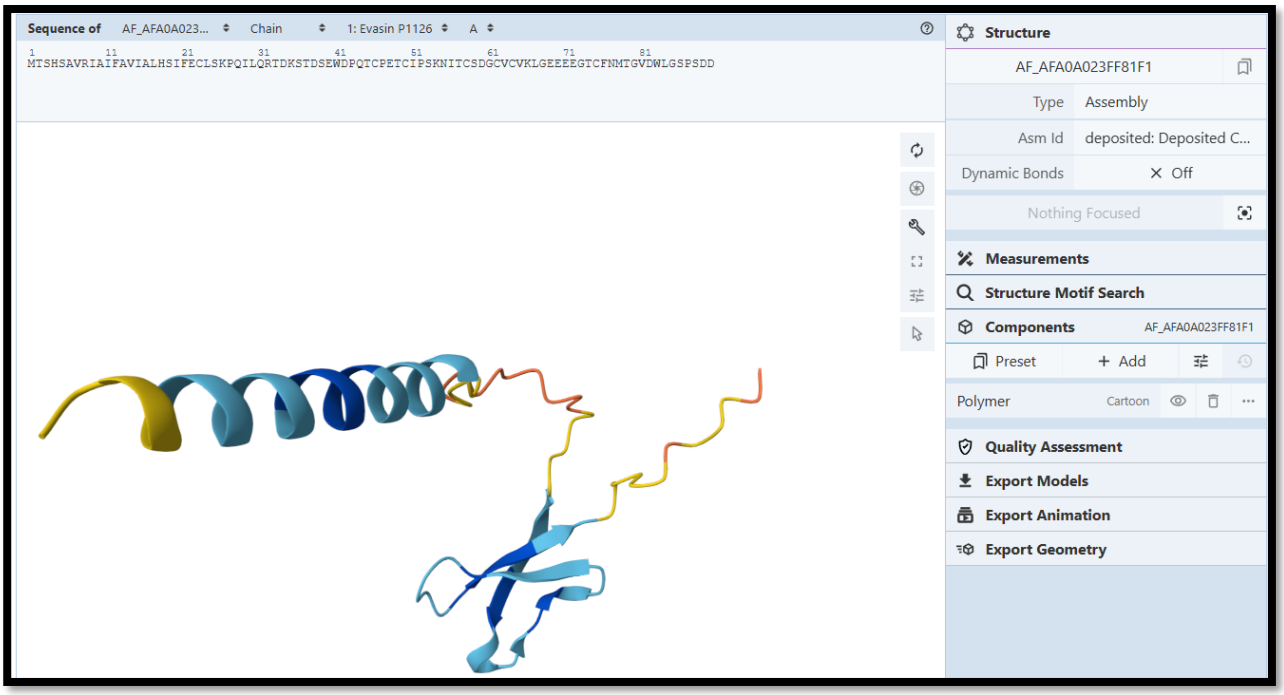


Fig. 13. Computed structure model of the Protein Evasin P1126

Fig. 13 presents an example of a protein structure predicted using AlphaFold and visualized through the AlphaFold Database interface. Such predicted structures provide valuable structural information for proteins lacking experimentally determined models and support downstream analyses including functional annotation, molecular docking, and protein engineering studies.

9 CONCLUSIONS

Protein structure prediction has evolved from traditional physics-based approaches to advanced DL-based frameworks. Modern systems now utilize end-to-end differentiable models that combine physical and biological constraints, significantly reducing the reliance on template matching. Notable DL models such as AlphaFold and RoseTTAFold offer near-experimental accuracy for globular proteins. This shift has led to more advanced architectures that leverage sequence data, inter-residue distance maps, and 3D coordinates through attention mechanisms. Innovations like ESMFold facilitate structure prediction from single sequences, circumventing Multiple Sequence Alignments (MSAs) and enhancing structural coverage at genomic scales. Furthermore, contemporary PSP increasingly merges computational predictions with limited experimental data, such as cryo-EM and mass spectrometry, to address complex and dynamic protein assemblies. The diffusion models and transformer-based architectures are being explored to capture dynamic conformational ensembles rather than static structures. Overall, recent advances in DL have transformed PSP into a highly accurate and practical computational tool, establishing a strong foundation for future developments in structural biology, drug discovery, and protein engineering.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

9.1. Future Directions

A few emerging research directions are expected to further improve the accuracy, efficiency, and applicability of PSP models.

i. Generative Methods and Diffusion Models
- Integration of diffusion models with physics-based energy functions.
- Design of proteins with specific biological functions (e.g. enzymes, antibodies).
- Combination of diffusion models with molecular dynamics simulations for the production of physically realistic ensembles of proteins.

ii. Foundation Models and Protein Language Models (PLMs)
The application of Large Language Models to biology assumes that amino acid sequences are a kind of natural language. Foundation models like ESM are trained on billions of sequences to learn "universal biological representations".
- Development of interactive AI assistants for biological reasoning and protein engineering.
- Natural-language-guided protein design where researchers describe a function and get a candidate sequence.
- Constructing trillion-parameter models that combine evolutionary, structural and functional information into a unified framework.

iii. Geometric Deep Learning
- Proteins are inherently three-dimensional biological macromolecules. Geometric DL employs specialized neural network architectures that preserve the spatial and symmetry properties of protein structures.
- Combining geometric DL with predicted flexibility measures (B-factors, pLDDT scores) and attention mechanisms.

iv. Multimodal Biological AI
- Future PSP systems will extend beyond sequence-only input to include a wide range of biological modalities. The "Multimodal transformer" approach aims to understand proteins in their larger cellular context
- Integrated Modalities consist of Protein structures and amino acid sequences, Gene expression data and "omics" data, Protein-protein interactions (PPI), Experimental dynamics (B-factors) and imaging in cells.
- The advantage of this method is a better analysis of disease related proteins, their behaviour and their interaction in living systems.

v. Structure-Function Co-Learning
- The ultimate frontier is the shift from predicting structure as an end point to predicting biological utility. So one end-to-end framework: Sequence → Structure → Dynamics → Function.
- Develop multi-task learning architectures to simultaneously predict 3D structure, B-factor/flexibility, binding sites and functional annotations from a single sequence input.

Although monomeric protein structure prediction has achieved high accuracy, significant challenges remain in modelling transient and heterogeneous protein–protein interactions and large macromolecular assemblies owing to limited experimental data. Beyond accuracy, the future of PSP must prioritize sustainability through the development of efficient, lightweight, and quantized architectures capable of real-time inference on modest hardware. As these tools become increasingly important in drug discovery and synthetic biology, improving model interpretability will become equally important. Explainable AI (XAI) frameworks can help elucidate the physical basis of predictions while robust biosecurity protocols can mitigate potential risks associated with dual-use protein design. Collectively, these research directions indicate that the next generation of PSP systems will extend beyond accurate structure prediction toward comprehensive, interpretable, and biologically informed models capable of supporting a wide range of scientific and biomedical applications.

FUNDING INFORMATION
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT
This study did not involve human or animal subjects and, therefore, did not require ethical approval.

STATEMENT OF CONFLICT OF INTERESTS
The authors declare no conflicts of interest related to this study.

REFERENCES

[1] Engelbert Buxbaum, *Fundamentals of protein structure and function*. Springer, 2015, doi: 10.1007/978-3-319-19920-7.

[2] X. Han, A. Sit, C. Christoffer, S. Chen, and D. Kihara, "A global map of the protein shape universe," *PLoS Computational Biology*, vol. 15, no. 4, p. e1006969, Apr. 2019, doi: 10.1371/journal.pcbi.1006969.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

[3] Z. Ding, N. Wang, N. Ji, and Z.-S. Chen, "Proteomics technologies for cancer liquid biopsies," *Molecular Cancer*, vol. 21, no. 1, p. 53, Feb. 2022, doi: 10.1186/s12943-022-01526-8.

[4] T. J. Hedl *et al.*, "Proteomics approaches for biomarker and drug target discovery in ALS and FTD," *Frontiers in Neuroscience*, vol. 13, p. 548, Jun. 2019, doi: 10.3389/fnins.2019.00548.

[5] R. Shelin and S. Meenakshi, "Rise of bacterial small proteins and peptides in therapeutic applications," *Protein and Peptide Letters*, vol. 30, no. 2, pp. 126–136, Jan. 2023, doi: 10.2174/0929866530666230118144723.

[6] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, no. 4096, pp. 223–230, Jul. 1973, doi: 10.1126/science.181.4096.223.

[7] S. Ahlawat, K. R. Mote, N.-A. Lakomek, and V. Agarwal, "Solid-State NMR: Methods for Biological solids," *Chemical Reviews*, vol. 122, no. 10, pp. 9643–9737, Mar. 2022, doi: 10.1021/acs.chemrev.1c00852.

[8] K. A. Dill and J. L. MacCallum, "The Protein-Folding Problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, Nov. 2012, doi: 10.1126/science.1219021.

[9] Daniel John Rigden, *From Protein Structure to Function with Bioinformatics*. Springer, 2009, doi:10.1007/978-1-4020-9058-5.

[10] R. Adiyaman and L. J. McGuffin, "Methods for the refinement of protein structure 3D models," *International Journal of Molecular Sciences*, vol. 20, no. 9, p. 2301, May 2019, doi: 10.3390/ijms20092301.

[11] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[12] W. L. Bragg, J. C. Kendrew, and M. F. Perutz, "Polypeptide chain configurations in crystalline proteins," *Proceedings of the Royal Society of London a Mathematical and Physical Sciences*, vol. 203, no. 1074, pp. 321–357, Oct. 1950, doi: 10.1098/rspa.1950.0142.

[13] V. Apostolopoulos *et al.*, "A Global Review on Short Peptides: Frontiers and Perspectives," *Molecules*, vol. 26, no. 2, p. 430, Jan. 2021, doi: 10.3390/molecules26020430.

[14] A. Vaseghi, M. Faridounnia, S. Shokrollahzade, S. Jahandideh, and K.-C. Chou, "Prediction of Protein Quaternary Structures," *Pattern Recognition in Computational Molecular Biology: Techniques and Approaches*, pp. 249–265, 2015, doi: 10.1002/9781119078845.ch14.

[15] M. Torrisi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1301–1310, Jan. 2020, doi: 10.1016/j.csbj.2019.12.011.

[16] R. P. D. Bank, "PDB Statistics: Overall growth of released structures per year." https://www.rcsb.org/stats/growth/growth-released-structures

[17] M. Dorn, M. B. E. Silva, L. S. Buriol, and L. C. Lamb, "Three-dimensional protein structure prediction: Methods and computational strategies," *Computational Biology and Chemistry*, vol. 53, pp. 251–276, Oct. 2014, doi: 10.1016/j.compbiolchem.2014.10.001.

[18] H. Deng, Y. Jia, and Y. Zhang, "Protein structure prediction," *International Journal of Modern Physics B*, vol. 32, no. 18, p. 1840009, Dec. 2017, doi: 10.1142/s021797921840009x.

[19] J. Lee, P. L. Freddolino, and Y. Zhang, "Ab initio protein structure prediction," in *From protein structure to function with bioinformatics*, pp. 3–35,2017 doi: 10.1007/978-94-024-1069-3_1.

[20] B. Kuhlman and P. Bradley, "Advances in protein structure prediction and design," *Nature Reviews Molecular Cell Biology*, vol. 20, no. 11, pp. 681–697, Aug. 2019, doi: 10.1038/s41580-019-0163-x.

[21] J. Won, M. Baek, B. Monastyrskyy, A. Kryshtafovych, and C. Seok, "Assessment of protein model structure accuracy estimation in CASP13: Challenges in the era of deep learning," *Proteins Structure Function and Bioinformatics*, vol. 87, no. 12, pp. 1351–1360, Aug. 2019, doi: 10.1002/prot.25804.

[22] B. Wen *et al.*, "Deep learning in proteomics," *PROTEOMICS*, vol. 20, no. 21–22, p. e1900335, Sep. 2020, doi: 10.1002/pmic.201900335.

[23] R. Pearce and Y. Zhang, "Deep learning techniques have significantly impacted protein structure prediction and protein design," *Current Opinion in Structural Biology*, vol. 68, pp. 194–207, Feb. 2021, doi: 10.1016/j.sbi.2021.01.007.

[24] E. Laine, S. Eismann, A. Elofsson, and S. Grudinin, "Protein sequence-to-structure learning: Is this the end(-to-end revolution)?," *Proteins Structure Function and Bioinformatics*, vol. 89, no. 12, pp. 1770–1786, Sep. 2021, doi: 10.1002/prot.26235.

[25] E. Callaway, "What's next for AlphaFold and the AI protein-folding revolution," *Nature*, vol. 604, no. 7905, pp. 234–238, Apr. 2022, doi: 10.1038/d41586-022-00997-5.

[26] B. Huang *et al.*, "Protein Structure Prediction: Challenges, advances, and the shift of research paradigms," *Genomics Proteomics & Bioinformatics*, vol. 21, no. 5, pp. 913–925, Mar. 2023, doi: 10.1016/j.gpb.2022.11.014.

[27] L. M. F. Bertoline, A. N. Lima, J. E. Krieger, and S. K. Teixeira, "Before and after AlphaFold2: An overview of protein structure prediction," *Frontiers in Bioinformatics*, vol. 3, p. 1120370, Feb. 2023, doi: 10.3389/fbinf.2023.1120370.

[28] A. Elofsson, "Progress at protein structure prediction, as seen in CASP15," *Current Opinion in Structural Biology*, vol. 80, p. 102594, Apr. 2023, doi: 10.1016/j.sbi.2023.102594.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

[29] F. Sapundzhi and V. Slavov, "RMSD calculations and computer modelling of protein structures," *Journal of Chemical Technology and Metallurgy*, vol. 55, no. 5, pp. 935–938, 2020. https://journal.uctm.edu/node/j2020-5/2_19-57_p935-938.pdf

[30] T. J. P. Hubbard, "RMS/coverage graphs: a qualitative method for comparing three-dimensional protein structure predictions," *Proteins: Structure, Function, and Bioinformatics*, vol. 37, no. S3, pp. 15–21, 1999, doi: 10.1002/(SICI)1097-0134(1999)37:3+<15::AID-PROT4>3.0.CO;2-Z.

[31] A. Zemla, "LGA: a method for finding 3D similarities in protein structures," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3370–3374, Jun. 2003, doi: 10.1093/nar/gkg571.

[32] Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins Structure Function and Bioinformatics*, vol. 57, no. 4, pp. 702–710, Oct. 2004, doi: 10.1002/prot.20264.

[33] I. Kufareva and R. Abagyan, "Methods of protein structure comparison," *Methods in Molecular Biology*, vol. 857, pp. 231–257, Jan. 2011, doi: 10.1007/978-1-61779-588-6_10.

[34] S. Kwon, J. Won, A. Kryshtafovych, and C. Seok, "Assessment of protein model structure accuracy estimation in CASP14: Old and new challenges," *Proteins Structure Function and Bioinformatics*, vol. 89, no. 12, pp. 1940–1948, Jul. 2021, doi: 10.1002/prot.26192.

[35] V. Mariani, M. Biasini, A. Barbato, and T. Schwede, "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests," *Bioinformatics*, vol. 29, no. 21, pp. 2722–2728, Aug. 2013, doi: 10.1093/bioinformatics/btt473.

[36] D. A. Case *et al.*, "Amber 10," 2008, University of California, San Francisco.

[37] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations," *Journal of Computational Chemistry*, vol. 4, no. 2, pp. 187–217, Jun. 1983, doi: 10.1002/jcc.540040211.

[38] K. F. Lau and K. A. Dill, "A lattice statistical mechanics model of the conformational and sequence spaces of proteins," *Macromolecules*, vol. 22, no. 10, pp. 3986–3997, Oct. 1989, doi: 10.1021/ma00200a030.

[39] N. London, B. Raveh, E. Cohen, G. Fathi, and O. Schueler-Furman, "Rosetta FlexPepDock web server—high resolution modeling of peptide–protein interactions," *Nucleic Acids Research*, vol. 39, no. suppl_2, pp. W249–W253, May 2011, doi: 10.1093/nar/gkr431.

[40] G. Van Houdt, C. Mosquera, and G. Nápoles, "A review on the long short-term memory model," *Artificial Intelligence Review*, vol. 53, no. 8, pp. 5929–5955, May 2020, doi: 10.1007/s10462-020-09838-1.

[41] S. M. Kandathil, J. G. Greener, and D. T. Jones, "Recent developments in deep learning applied to protein structure prediction," *Proteins Structure Function and Bioinformatics*, vol. 87, no. 12, pp. 1179–1189, Oct. 2019, doi: 10.1002/prot.25824.

[42] J. Eickholt and J. Cheng, "A study and benchmark of DNcon: a method for protein residue-residue contact prediction using deep networks," *BMC Bioinformatics*, vol. 14, no. S14, p. S12, Oct. 2013, doi: 10.1186/1471-2105-14-s14-s12.

[43] S. Seemayer, M. Gruber, and J. Söding, "CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations," *Bioinformatics*, vol. 30, no. 21, pp. 3128–3130, Jul. 2014, doi: 10.1093/bioinformatics/btu500.

[44] D. T. Jones, T. Singh, T. Kosciolek, and S. Tetchner, "MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins," *Bioinformatics*, vol. 31, no. 7, pp. 999–1006, Nov. 2014, doi: 10.1093/bioinformatics/btu791.

[45] M. J. Skwark, A. Abdel-Rehim, and A. Elofsson, "PconsC: combination of direct information methods and alignments improves contact prediction," *Bioinformatics*, vol. 29, no. 14, pp. 1815–1816, May 2013, doi: 10.1093/bioinformatics/btt259.

[46] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell, "Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models," *Physical Review E*, vol. 87, no. 1, p. 012707, Jan. 2013, doi: 10.1103/physreve.87.012707.

[47] S. M. Mortuza, W. Zheng, C. Zhang, Y. Li, R. Pearce, and Y. Zhang, "Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions," *Nature Communications*, vol. 12, no. 1, p. 5011, Aug. 2021, doi: 10.1038/s41467-021-25316-w.

[48] 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 Computational Biology*, vol. 13, no. 1, p. e1005324, Jan. 2017, doi: 10.1371/journal.pcbi.1005324.

[49] F. Ju *et al.*, "CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction," *Nature Communications*, vol. 12, no. 1, p. 2535, May 2021, doi: 10.1038/s41467-021-22869-8.

[50] Z. Du *et al.*, "The trRosetta server for fast and accurate protein structure prediction," *Nature Protocols*, vol. 16, no. 12, pp. 5634–5651, Nov. 2021, doi: 10.1038/s41596-021-00628-9.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

[51] J. Liu, G.-X. He, K.-L. Zhao, and G.-J. Zhang, "De novo protein structure prediction by incremental inter-residue geometries prediction and model quality assessment using deep learning," *bioRxiv (Cold Spring Harbor Laboratory)*, Jan. 2022, doi: 10.1101/2022.01.11.475831.

[52] Y. Li, C. Zhang, D.-J. Yu, and Y. Zhang, "Deep learning geometrical potential for high-accuracy ab initio protein structure prediction," *iScience*, vol. 25, no. 6, p. 104425, May 2022, doi: 10.1016/j.isci.2022.104425.

[53] M. AlQuraishi, "End-to-End differentiable learning of protein structure," *Cell Systems*, vol. 8, no. 4, pp. 292-301.e3, Apr. 2019, doi: 10.1016/j.cels.2019.03.006.

[54] J. Ingraham, A. Riesselman, C. Sander, and D. Marks, "Learning Protein Structure with a Differentiable Simulator," *International conference on learning representations, 2019*. https://openreview.net/forum?id=Byg3y3C9Km

[55] A. W Senior *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, Jan. 2020, doi: 10.1038/s41586-019-1923-7.

[56] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Jul. 2021, doi: 10.1038/s41586-021-03819-2.

[57] M. Baek *et al.*, "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, Jul. 2021, doi: 10.1126/science.abj8754.

[58] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, and M. Steinegger, "ColabFold: making protein folding accessible to all," *Nature Methods*, vol. 19, no. 6, pp. 679–682, May 2022, doi: 10.1038/s41592-022-01488-1.

[59] S. Cheng *et al.*, "FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours," *arXiv.org*, Mar. 02, 2022. https://arxiv.org/abs/2203.00854

[60] Z. Li *et al.*, "Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold," *bioRxiv (Cold Spring Harbor Laboratory)*, Aug. 2022, doi: 10.1101/2022.08.04.502811.

[61] J. Zhang *et al.*, "Unsupervisedly prompting AlphaFold2 for accurate Few-Shot protein structure prediction," *Journal of Chemical Theory and Computation*, vol. 19, no. 22, pp. 8460–8471, Nov. 2023, doi: 10.1021/acs.jctc.3c00528.

[62] G. Wang *et al.*, "HelixFold: An Efficient Implementation of AlphaFold2 using PaddlePaddle," *arXiv.org*, Jul. 12, 2022. https://arxiv.org/abs/2207.05477

[63] G. Ahdritz *et al.*, "OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization," *Nature Methods*, vol. 21, no. 8, pp. 1514–1524, May 2024, doi: 10.1038/s41592-024-02272-z.

[64] T. Gong, F. Ju, S. Sun and D. Bu, "SASA-Net: A Spatial-Aware Self-Attention Mechanism for Building Protein 3D Structure Directly From Inter- Residue Distances," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 6, pp. 3482-3488, Nov.-Dec. 2023, doi: 10.1109/TCBB.2023.3240456.

[65] W. Zhong and F. Gu, "Predicting Local Protein 3D Structures Using Clustering Deep Recurrent Neural Network," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 1, pp. 593-604, 1 Jan.-Feb. 2022, doi: 10.1109/TCBB.2020.3005972.

[66] F. F. Alam and A. Shehu, "Variational Autoencoders for Protein Structure Prediction," in *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '20)*, Virtual Event, USA, 2020, Art. no. 27, pp. 1–10, doi: 10.1145/3388440.3412471.

[67] X. Guo, Y. Du, S. Tadepalli, L. Zhao, and A. Shehu, "Generating tertiary protein structures via interpretable graph variational autoencoders," *Bioinformatics Advances*, vol. 1, no. 1, p. vbab036, Jan. 2021, doi: 10.1093/bioadv/vbab036.

[68] R. Rao, J. Meier, T. Sercu, S. Ovchinnikov, and A. Rives, "Transformer protein language models are unsupervised structure learners," *bioRxiv (Cold Spring Harbor Laboratory)*, Dec. 2020, doi: 10.1101/2020.12.15.422761.

[69] A. Rives *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, Apr. 2021, doi: 10.1073/pnas.2016239118.

[70] J. H. Lee *et al.*, "EquiFold: Protein Structure Prediction with a Novel Coarse-Grained Structure Representation," *bioRxiv (Cold Spring Harbor Laboratory)*, Oct. 2022, doi: 10.1101/2022.10.07.511322.

[71] T. Bepler and B. Berger, "Learning the protein language: Evolution, structure, and function," *Cell Systems*, vol. 12, no. 6, pp. 654-669.e3, Jun. 2021, doi: 10.1016/j.cels.2021.05.017.

[72] R. Chowdhury *et al.*, "Single-sequence protein structure prediction using a language model and deep learning," *Nature Biotechnology*, vol. 40, no. 11, pp. 1617–1623, Oct. 2022, doi: 10.1038/s41587-022-01432-w.

[73] Y. Hong, J. Lee, and J. Ko, "A-Prot: protein structure modeling using MSA transformer," *BMC Bioinformatics*, vol. 23, no. 1, p. 93, Mar. 2022, doi: 10.1186/s12859-022-04628-8.

[74] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic level protein structure with a language model," *bioRxiv (Cold Spring Harbor Laboratory)*, Jul. 2022, doi: 10.1101/2022.07.20.500902.

[75] R. Wu *et al.*, "High-resolution de novo structure prediction from primary sequence," *bioRxiv (Cold Spring Harbor Laboratory)*, Jul. 2022, doi: 10.1101/2022.07.21.500999.

[76] Z. Lin *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, Mar. 2023, doi: 10.1126/science.ade2574.

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

[77] X. Fang *et al.*, "A method for multiple-sequence-alignment-free protein structure prediction using a protein language model," *Nature Machine Intelligence*, vol. 5, no. 10, pp. 1087–1096, Oct. 2023, doi: 10.1038/s42256-023-00721-6.

[78] T. D. Barrett *et al.*, "So ManyFolds, So Little Time: Efficient protein structure prediction with PLMS and MSAs," *bioRxiv (Cold Spring Harbor Laboratory)*, Oct. 2022, doi: 10.1101/2022.10.15.511553.

[79] X. Jing, F. Wu, X. Luo, and J. Xu, "Single-sequence protein structure prediction by integrating protein language models," *Proceedings of the National Academy of Sciences*, vol. 121, no. 13, p. e2308788121, Mar. 2024, doi: 10.1073/pnas.2308788121.

[80] B. Adhikari and J. Cheng, "Protein residue contacts and prediction methods," *Methods in Molecular Biology*, vol. 1415, pp. 463–476, Jan. 2016, doi: 10.1007/978-1-4939-3572-7_24.

[81] J. Skolnick and M. Gao, "The role of local versus nonlocal physicochemical restraints in determining protein native structure," *Current Opinion in Structural Biology*, vol. 68, pp. 1–8, Oct. 2020, doi: 10.1016/j.sbi.2020.10.008.

[82] H. Zhang and Y. Shen, "Template-based prediction of protein structure with deep learning," *BMC Genomics*, vol. 21, no. S11, p. 878, Dec. 2020, doi: 10.1186/s12864-020-07249-8.

[83] G. Qi, M. R. Tollefson, R. A. Gogal, R. J. H. Smith, M. AlQuraishi, and M. J. Schnieders, "Protein structure prediction using a maximum likelihood formulation of a recurrent geometric network," *bioRxiv (Cold Spring Harbor Laboratory)*, Sep. 2021, doi: 10.1101/2021.09.03.458873.

[84] V. Nallasamy and M. Seshiah, "Protein structure prediction using quantile dragonfly and structural Class-Based deep learning," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 36, no. 04, Mar. 2022, doi: 10.1142/s021800142250015x.

[85] F. Marchi and R. S. Parpinelli, "A Multi-objective Approach to the Protein Structure Prediction Problem using the Biased Random-Key Genetic Algorithm," *2021 IEEE Congress on Evolutionary Computation (CEC)*, Kraków, Poland, 2021, pp. 1070-1077, doi: 10.1109/CEC45853.2021.9504745.

[86] V. Nallasamy and M. Seshiah, "Bartletts principal regressive and arbitrary African buffalo optimizatized three-dimensional protein structure prediction," *Artificial Intelligence Review*, vol. 57, no. 1, Jan. 2024, doi: 10.1007/s10462-023-10634-w.

[87] S. M. S. Rayesha, W. A. Banu, and S. Priya, "The prediction of protein structure using neural network," in *Lecture notes in networks and systems*, 2023, pp. 1021–1028. doi: 10.1007/978-981-99-1414-2_69.

[88] Y. Zhou, K. Tan, X. Shen, Z. He and H. Zheng, "A Protein Structure Prediction Approach Leveraging Transformer and CNN Integration," *2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE)*, Shanghai, China, 2024, pp. 749-753, doi: 10.1109/ICAACE61206.2024.10548253.

[89] C. S. Srushti, P. P.M, V. A and K. R. Venugopal, "Prediction and Analysis of Protein 3D Structures Using Protein Language Model and Streamlit," *2024 International Conference on Information Management and Technology (ICIMTech)*, Bali, Indonesia, 2024, pp. 636-641, doi: 10.1109/ICIMTech63123.2024.10780906.

[90] J. Hu, W. Wang, and H. Gong, "Cerebra: a computationally efficient framework for accurate protein structure prediction," *bioRxiv (Cold Spring Harbor Laboratory)*, Feb. 2024, doi: 10.1101/2024.02.02.578551.

[91] T. Xie, Z. Song, and J. Huang, "Conditioned Protein Structure prediction," *PRX Life*, vol. 2, no. 4, Oct. 2024, doi: 10.1103/prxlife.2.043001.

[92] B. A. A. Team *et al.*, "Protenix - advancing structure prediction through a comprehensive AlphaFold3 reproduction," *bioRxiv (Cold Spring Harbor Laboratory)*, Jan. 2025, doi: 10.1101/2025.01.08.631967.

[93] Y. Gao, H. Wang, J. Zhou, and Y. Yang, "An easy-to-use three-dimensional protein-structure-prediction online platform 'DPL3D' based on deep learning algorithms," *Current Research in Structural Biology*, vol. 9, p. 100163, Jan. 2025, doi: 10.1016/j.crstbi.2024.100163.

[94] R. S. Silva and R. S. Parpinelli, "A Self-adaptive Differential Evolution with Fragment Insertion for the Protein Structure Prediction Problem," in *Lecture notes in computer science*, 2018, pp. 136–149. doi: 10.1007/978-3-030-05983-5_10.

[95] M. Boiani and R. S. Parpinelli, "A GPU-based hybrid jDE algorithm applied to the 3D-AB protein structure prediction," *Swarm and Evolutionary Computation*, vol. 58, p. 100711, Jun. 2020, doi: 10.1016/j.swevo.2020.100711.

[96] R. S. Parpinelli, N. N. Will, and R. S. Da Silva, "A self-adaptive evolutionary algorithm using Monte Carlo Fragment insertion and conformation clustering for the protein structure prediction problem," *Natural Computing*, vol. 23, no. 4, pp. 687–705, Aug. 2022, doi: 10.1007/s11047-022-09916-z.

[97] V. Nallasamy and M. S, "Bingham deep neural and oppositional fish swarm optimized protein structure prediction," *Journal of Biomolecular Structure and Dynamics*, vol. 40, no. 19, pp. 8706–8724, May 2021, doi: 10.1080/07391102.2021.1915181.

[98] P. Rachitskii, I. Kruglov, A. V. Finkelstein, and A. R. Oganov, "Protein structure prediction using the evolutionary algorithm USPEX," *Proteins Structure Function and Bioinformatics*, vol. 91, no. 7, pp. 933–943, Feb. 2023, doi: 10.1002/prot.26478.

[99] R. S. Parpinelli, N. Wojeicchowski and N. N. Will, "Protein Structure Prediction Using Dynamic Speciation Evolutionary Algorithm with Aggregated Problem Information," *2024 IEEE Conference on Computational*

International Journal of Emerging Research in Engineering, Science, and Management
Vol. 5 | Issue 2 | pp. 119-147 | Apr-Jun 2026.
www.ijeresm.com | eISSN – 2583-4894
Received: 30 Mar 2026 | Revised: 02 Jun 2026 | Accepted: 08 Jun 2026 | Published: 21 Jun 2026

Intelligence in Bioinformatics and Computational Biology (CIBCB), Natal, Brazil, 2024, pp. 1-8, doi: 10.1109/CIBCB58642.2024.10702172.

[100] B. Jing, E. Erives, P. Pao-Huang, G. Corso, B. Berger, and T. Jaakkola, "EigenFold: Generative Protein Structure Prediction with Diffusion Models," *arXiv.org*, Apr. 05, 2023. https://arxiv.org/abs/2304.02198

[101] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with AlphaFold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, May 2024, doi: 10.1038/s41586-024-07487-w.

[102] J. Wohlwend *et al.*, "Boltz-1 Democratizing Biomolecular Interaction Modeling," *bioRxiv (Cold Spring Harbor Laboratory)*, Nov. 2024, doi: 10.1101/2024.11.19.624167.

[103] C. Discovery *et al.*, "Chai-1: Decoding the molecular interactions of life," *bioRxiv (Cold Spring Harbor Laboratory)*, Oct. 2024, doi: 10.1101/2024.10.10.615955.

[104] M. Varadi *et al.*, "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, vol. 50, no. D1, pp. D439–D444, Oct. 2021, doi: 10.1093/nar/gkab1061.

[105] R. Tian *et al.*, "A Pharmacoinformatics Analysis of Artemisinin Targets and de novo Design of Hits for Treating Ulcerative Colitis," *Frontiers in Pharmacology*, vol. 13, p. 843043, Mar. 2022, doi: 10.3389/fphar.2022.843043.

[106] G. Kobakhidze, A. Sethi, S. Valimehr, S. A. Ralph, and I. Rouiller, "The AAA+ ATPase p97 as a novel parasite and tuberculosis drug target," *Trends in Parasitology*, vol. 38, no. 7, pp. 572–590, Apr. 2022, doi: 10.1016/j.pt.2022.03.004.

[107] P.-S. Huang, S. E. Boyken, and D. Baker, "The coming of age of de novo protein design," *Nature*, vol. 537, no. 7620, pp. 320–327, Sep. 2016, doi: 10.1038/nature19946.

[108] P. Bryant, G. Pozzati, and A. Elofsson, "Improved prediction of protein-protein interactions using AlphaFold2," *Nature Communications*, vol. 13, no. 1, p. 1265, Mar. 2022, doi: 10.1038/s41467-022-28865-w.