molecules

Review

Recent Progress of Protein Tertiary Structure Prediction

Qiqige Wuyun [1,*], Yihan Chen [2], Yifeng Shen [3], Yang Cao [4,*], Gang Hu [5,*], Wei Cui [2,*], Jianzhao Gao [2,*] and Wei Zheng [6,*]

[1] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA
[2] School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China; yh_chen@mail.nankai.edu.cn
[3] Faculty of Environment and Information Studies, Keio University, Fujisawa 252-0882, Kanagawa, Japan; tomshen@keio.jp
[4] College of Life Sciences, Sichuan University, Chengdu 610065, China
[5] NITFID, School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin 300071, China
[6] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
[*] Correspondence: wuyunqiq@msu.edu (Q.W.); cao@scu.edu.cn (Y.C.); huggs@nankai.edu.cn (G.H.); weicui@nankai.edu.cn (W.C.); gaojz@nankai.edu.cn (J.G.); zhengwei@umich.edu (W.Z.); Tel.: +1-734-802-9414 (W.Z.)

Abstract: The prediction of three-dimensional (3D) protein structure from amino acid sequences has stood as a significant challenge in computational and structural bioinformatics for decades. Recently, the widespread integration of artificial intelligence (AI) algorithms has substantially expedited advancements in protein structure prediction, yielding numerous significant milestones. In particular, the end-to-end deep learning method AlphaFold2 has facilitated the rise of structure prediction performance to new heights, regularly competitive with experimental structures in the 14th Critical Assessment of Protein Structure Prediction (CASP14). To provide a comprehensive understanding and guide future research in the field of protein structure prediction for researchers, this review describes various methodologies, assessments, and databases in protein structure prediction, including traditionally used protein structure prediction methods, such as template-based modeling (TBM) and template-free modeling (FM) approaches; recently developed deep learning-based methods, such as contact/distance-guided methods, end-to-end folding methods, and protein language model (PLM)-based methods; multi-domain protein structure prediction methods; the CASP experiments and related assessments; and the recently released AlphaFold Protein Structure Database (AlphaFold DB). We discuss their advantages, disadvantages, and application scopes, aiming to provide researchers with insights through which to understand the limitations, contexts, and effective selections of protein structure prediction methods in protein-related fields.

Keywords: AlphaFold2; contact map; deep learning; distance map; end-to-end methods; multi-domain proteins; protein language model; protein tertiary structure prediction; template-based modeling; template-free modeling

Citation: Wuyun, Q.; Chen, Y.; Shen, Y.; Cao, Y.; Hu, G.; Cui, W.; Gao, J.; Zheng, W. Recent Progress of Protein Tertiary Structure Prediction. *Molecules* **2024**, *29*, 832. https://doi.org/10.3390/molecules29040832

Academic Editor: Tibor Páli

Received: 30 December 2023
Revised: 6 February 2024
Accepted: 8 February 2024
Published: 13 February 2024

1. Introduction

Proteins are macromolecules that play important roles in facilitating the essential functions vital for life's sustenance. Their pivotal involvement spans a diverse array—providing structural support to cells, safeguarding the immune system, catalyzing crucial enzymatic reactions, orchestrating cellular signal transmission, regulating the intricate processes of transcription and translation, and encompassing the synthesis and breakdown of biomolecules. Moreover, they contribute significantly to the regulation of developmental processes, biological pathways, and the constitution of protein complexes and subcellular structures. These diverse and remarkable functions originate from their distinct three-dimensional (3D) structures, which vary across different protein molecules. Since Anfinsen showed that the tertiary structure of a protein is determined by its amino acid sequence in 1973 [1], understanding the protein sequence–structure–function paradigm has emerged as a fundamental cornerstone within modern biomedical studies. Due to significant efforts in genome sequencing over the last few decades [2–4], the number of known amino acid sequences deposited in UniProt [5] has grown to over 250 million. Despite the impressive number of data, the amino acid sequences themselves only offer limited insights into the biological functions of individual proteins, as these functions are primarily determined by their three-dimensional structures.

Some of the most widely used experimental techniques for determining protein structures include X-ray crystallography [6], NMR spectroscopy [7], and cryo-electron microscopy [8]. Despite their accuracy, the considerable human involvement and substantial expenses involved in experimentally resolving a protein's structure have hindered advancement in the number of solved protein structures. Consequently, the expansion in solved protein structures has considerably trailed the accumulation of protein sequences. At present, the Protein Data Bank [9] (PDB) contains structures for approximately 0.21 million proteins, accounting for less than 0.1% of the total sequences cataloged in the UniProt database [10]. This disparity highlights the ever-widening gap between known protein sequences and experimentally solved protein structures. Nevertheless, owing to substantial collective efforts within the scientific community in recent decades [11–25], computational approaches have made remarkable progress, through which an increasing fraction of sequences in various organisms have had their tertiary structures reliably modeled [26–39]. For example, the first version of AlphaFold demonstrated exceptional predictive capabilities in protein structure prediction by employing the deep learning-based distance map prediction during the 13th Critical Assessment of Protein Structure Prediction (CASP13). Furthermore, with the utilization of the end-to-end deep learning approach, the AlphaFold2 has facilitated the rise of structure prediction performance to new heights, regularly competitive with experimental structures in CASP14. These methodologies have significantly contributed to diverse biomedical investigations, including structure-based protein function annotation [40–44], mutation analysis [45–52], ligand screening [53–59], and drug discovery [60–65].

In this review, we start with an overview of the history of protein structure prediction, including template-based modeling (TBM) and template-free modeling (FM) methods. TBM techniques predict models by refining the structures of existing proteins, known as templates, identified from the PDB. In contrast, FM methods construct protein structures without relying on template structures. Then, we discuss the recent advancements and progress brought about by deep learning technologies, including contact/distance-guided protein structure prediction methods, end-to-end folding methods, and protein language model (PLM)-based methods. In particular, we highlight the breakthrough in end-to-end methods and protein language model (PLM)-based methods. Additionally, we introduce recent progress in multi-domain protein structure predictions. Finally, we describe the CASP experiments and some widely used assessment measures for protein structure prediction, followed by the introduction of the recently released AlphaFold Protein Structure Database (AlphaFold DB) and its corresponding applications.

Tables S1–S7 offer links to the methods discussed in this review, serving as a supplemental resource for readers' accessibility. Meanwhile, Figure 1 presents a comprehensive timeline of these methods and some significant achievements covered in this review.

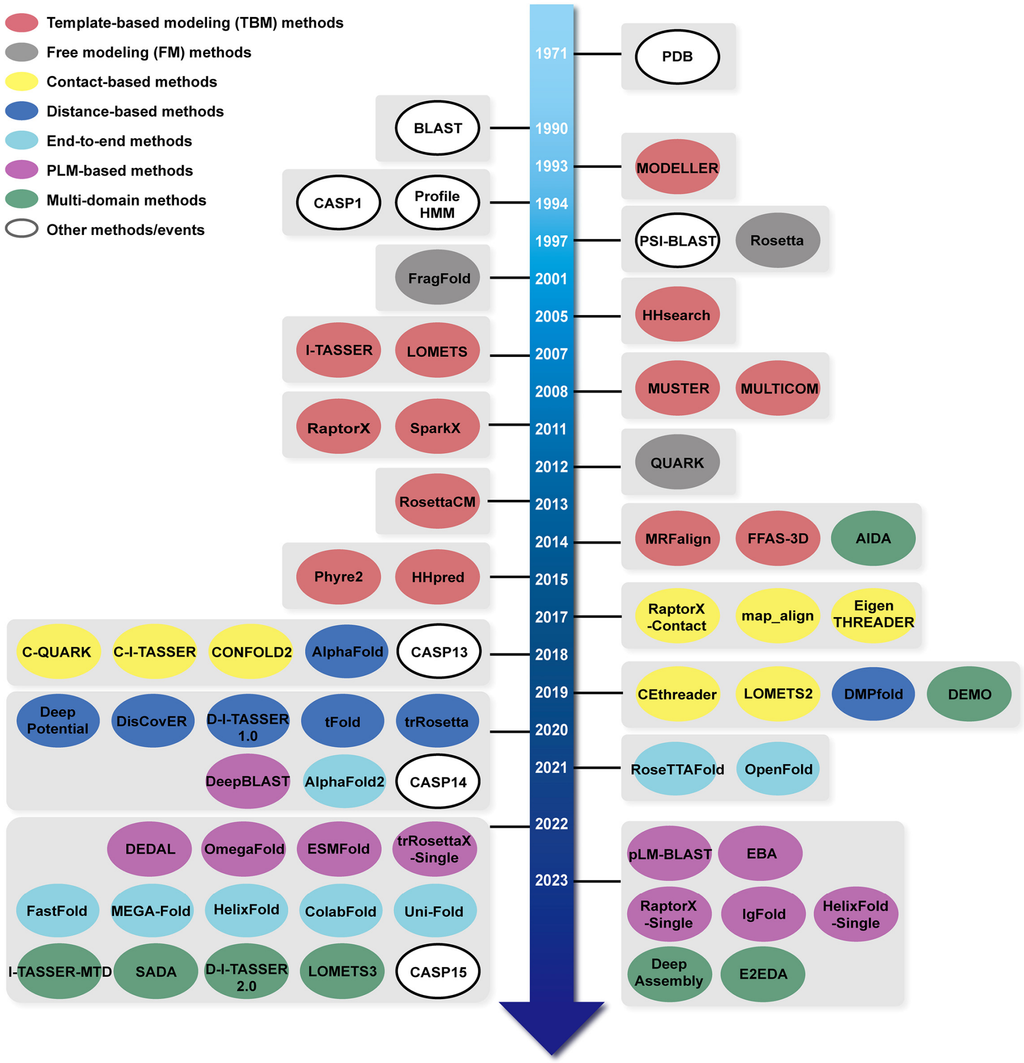

Figure 1. The timeline of important methods or tools in protein structure prediction. Different methods or tools are denoted by different colors: template-based modelling (TBM) methods are represented by red, free modeling (FM) methods by gray, contact-based methods by yellow, distance-based methods by blue, end-to-end-based methods by cyan, protein language model (PLM)-based methods by purple, and multi-domain methods by green, while other important methods or events are highlighted in white. Note that some methods may be categorized under two or more groups, but we only highlighted the most important category for each method.

2. An Overview of Protein Structure Prediction

2.1. *Template-Based Modeling (TBM) Methods*

Template-based modeling (TBM) methods have emerged as pivotal approaches in the realm of computational biology for predicting protein structures. TBM leverages known protein structures, referred to as templates, from the PDB to predict the structure of an unknown protein (target), assuming that the target shares a significant degree of sequence similarity with the template. As shown in Figure 2, TBM methods usually consist of the following four steps: (i) identifying templates related to the protein of interest, (ii) aligning the query protein with the templates, (iii) building the initial structural framework by replicating the aligned regions, and (iv) constructing the unaligned regions and refining the structure. TBM can be classified as homology modeling (comparative modeling), which is often employed when there is substantial sequence identity—typically 30% or greater—between the template and the protein of interest, and threading (fold recognition), which is used when the sequence identity drops below the 30% threshold [66].

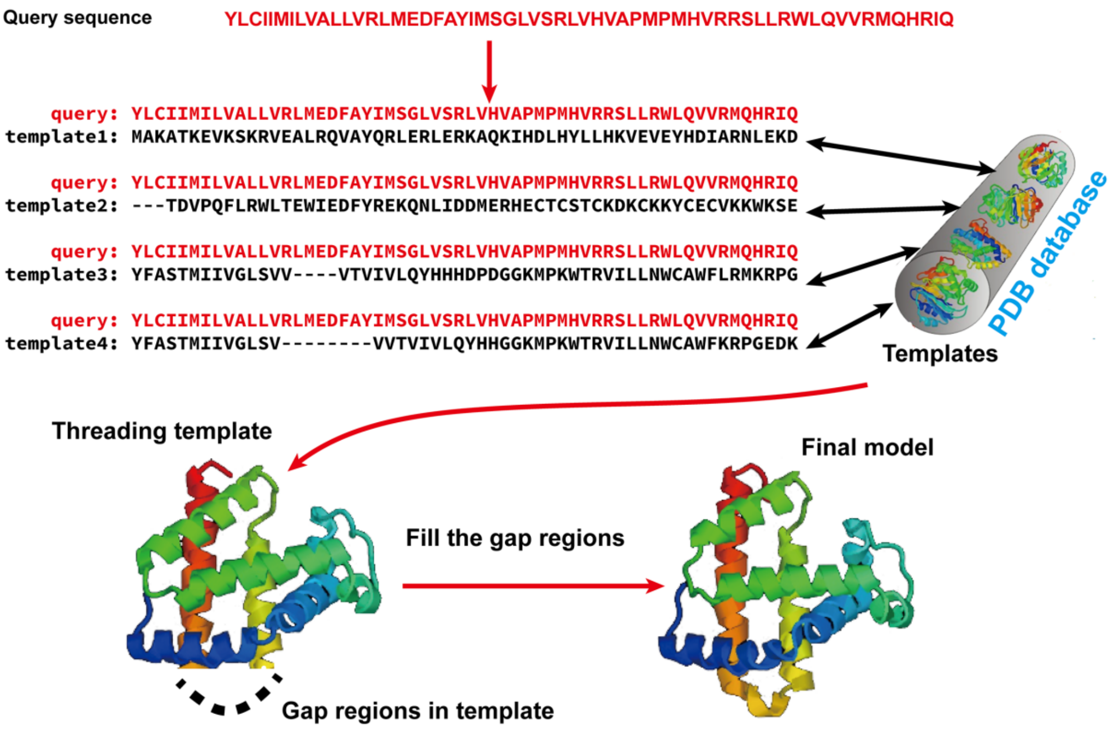
Query sequence
YLCIIMILVALLVRLMEDFAYIMSGLVSRLVHVAPMPMHVRRSLLRWLQVVRMQHRIQ
query: YLCIIMILVALLVRLMEDFAYIMSGLVSRLVHVAPMPMHVRRSLLRWLQVVRMQHRIQ
template1: MAKATKEVKSKRVEALRQVAYQRLERLERKAQKIHDLHYLLHKVEVEYHDIARNLEKD
query: YLCIIMILVALLVRLMEDFAYIMSGLVSRLVHVAPMPMHVRRSLLRWLQVVRMQHRIQ
template2: ---TDVPQFLRWLTEWIEDFYREKQNLIDDMERHECTCSTCKDKCKKYCECVKKWKSE
query: YLCIIMILVALLVRLMEDFAYIMSGLVSRLVHVAPMPMHVRRSLLRWLQVVRMQHRIQ
template3: YFASTMIIVGLSVV----VTVIVLQYHHHDPDGGKMPKWTRVILLNWCAWFLRMKRPG
query: YLCIIMILVALLVRLMEDFAYIMSGLVSRLVHVAPMPMHVRRSLLRWLQVVRMQHRIQ
template4: YFASTMIIVGLSV--------VVTVIVLQYHHGGKMPKWTRVILLNWCAWFKRPGEDK
PDB database
Templates
Threading template
Final model
Fill the gap regions
Gap regions in template

Figure 2. Illustration of template-based modeling (TBM) methods. Starting from a query sequence, templates are identified from Protein Data Bank (PDB) and subsequently aligned with the query protein sequence. Then, the final structural model is constructed by replicating the aligned regions and refining the unaligned regions.

In homology modeling, high-quality templates are detected and aligned using straightforward sequence–sequence alignment algorithms, such as dynamic programming-based techniques like the Needleman–Wunsch [67] algorithm for global alignment and the Smith–Waterman [68] algorithm for local alignment. BLAST [69] is another widely used tool to identify templates and generate alignments, which initially identified short matches between the query and template, and then extended these matches to generate alignments.

In threading, since the sequence identity between the best available template and the query protein falls below 30%, it is hard to identify templates simply based on straightforward sequence–sequence alignment algorithms. Hence, the 1D profile of local structural features is used to represent a template's 3D structure, because they are often more conserved than the amino acid identities themselves and, thus, can be used to identify and align proteins with similar structures but more distant sequence homology. A commonly used sequence profile is the Position-specific Scoring Matrix (PSSM), which captures the amino acid tendencies at each position within the multiple sequence alignment (MSA). The PSSM is iteratively employed to search through a template database, aiming to identify distantly homologous templates for a specific protein sequence. One popularly used profile-based threading algorithm is MUSTER [70], which combines various sequence and structural information into single-body terms in a dynamic programming search, as follows: (i) sequence profiles; (ii) secondary structures; (iii) structure fragment profiles; (iv) solvent accessibility; (v) dihedral torsion angles; and (vi) hydrophobic scoring matrix. In addition to PSSMs, profile hidden Markov models (HMMs) are another type of sequence profile. A profile HMM is a probabilistic model that captures the evolutionary changes within an MSA. The key advantage of profile HMMs lies in their utilization of position-specific gap penalties and substitution probabilities, providing a closer representation of the true underlying sequence distribution [71]. HHsearch [72] is the most widely used profile HMM-based threading method, which generalized the alignment of protein sequences with a profile HMM to the case of pairwise alignment of profile HMMs for detecting distant homologous relationships between proteins.

Given the recent substantial improvements in contact and distance map prediction using deep learning, which will be discussed later, threading methods guided by these maps represent the cutting edge in fold recognition, achieving superior accuracy compared to general profile or profile HMM-based threading methods. Among these approaches, EigenTHREADER [73] utilized the eigen decomposition of contact maps to derive the primary eigenvectors, which were used for aligning the template and query contact maps. CEthreader [74], employing a similar eigen decomposition strategy, outperformed pure contact map-based threading methods by integrating data from local structural feature prediction and sequence-based profiles. map_align [21], on the other hand, introduced an iterative dual dynamic programming technique to align contact maps, while DeepThreader [75] leveraged predicted distance maps to establish alignments. Most recently, DisCovER [76] integrated deep learning-predicted distance and orientation into the threading method by generating alignments through an iterative double dynamic programming framework. In addition, meta-threading approaches, such as LOMETS [77–79], combine the templates' output, via multiple threading programs, into a set of consensus templates, thereby attaining enhanced accuracy. For example, LOMETS2 [78] integrated a comprehensive set of state-of-the-art threading programs, including contact-guided threading approaches, and utilizes deep profiles generated by a novel deep MSA construction method, DeepMSA [80].

Furthermore, deep learning-based methods have been directly applied to recognize distant homology templates. The cutting-edge methods, such as ThreaderAI [81] and SAdLSA [82], conceptualize the task of aligning query sequence with template as the classical pixel classification problem in computer vision, which allows for the integration of a deep residual neural network [83] into fold recognition. More recently, the application of language models, originally developed for text classification and generative tasks, to protein sequences marks a significant advancement in the bioinformatics field. Protein language models (PLMs) are a type of neural network with self-supervised training on an extensive number of protein sequences [84,85]. Once trained, PLMs can be used to rapidly generate high-dimensional embeddings on a per-residue level, which can be viewed as a "semantic meaning" of each amino acid within the context of the full protein sequence. Such representations have proven invaluable in identifying distant homologous relationships between proteins. For example, pLM-BLAST [86] detected distant homologous relationships by integrating single-sequence embeddings, obtained from protein language models (PLMs), with a local similarity detection algorithm from BLAST. pLM-BLAST operated on an unsupervised basis, eliminating the need for training a specialized deep-learning model, and was capable of computing both local and global alignments, leveraging the strengths of PLM-derived embeddings and BLAST-based algorithms. EBA [87] was a new tool designed to generate embedding-based protein sequence alignments, particularly in the challenging 'twilight zone'. It leveraged the distances between all possible pairs of residue embeddings to create a "similarity matrix." This matrix subsequently served as a scoring matrix within a classical dynamic programming alignment framework. The absence of any requirement for training and parameter optimization, coupled with its flexibility to any language model, rendered the EBA method robust to generalization and easy to interpret. DEDAL [88] and DeepBLAST [89] both integrated residue embeddings learned from a PLM into a differentiable alignment framework; however, DEDAL used an affine scoring function, while DeepBLAST had a simpler linear model for scores and only produced global alignments. Due to their rich information contents, sequence embeddings produced by PLMs have been successfully applied to many other tasks, especially in the prediction of tertiary structures, which will be discussed later.

Once the templates are identified and aligned with the query proteins, the subsequent step involves building a model by replicating and refining the structure of the template. The most widely used method was MODELLER [16], which constructed tertiary structure models by optimally satisfying spatial constraints extracted from the template alignments, along with other general structural constraints, such as ideal bond lengths, bond angles, and dihedral angles. Furthermore, the new HHpred modeling pipeline, proposed by the Söding group, has extended the MODELLER by employing (i) atomic distance restraints described by two-component Gaussian mixtures, (ii) optimal weights to correct for redundancy among related templates, and (iii) a heuristic template selection strategy [90].

With the development of computational techniques, some methods are proposed to convert alignments directly into 3D models. A notable example is I-TASSER [91–93], an extension of TASSER [28]. This method utilized a process wherein continuous fragments were extracted from the aligned regions of multiple threading templates identified by LOMETS. These fragments were reassembled during structure assembly simulations. I-TASSER incorporated constraints derived from template alignments and a set of knowledge-based energy terms. These energy terms included hydrogen bonding, secondary structure formation, and side-chain contact formation. The integration of these components was used to guide the Replica Exchange Monte Carlo (REMC) simulation. After clustering low-energy decoys and selecting the centroid of the most favorable cluster, the centroid was compared against the PDB to identify additional templates. The constraints from these new templates, combined with those from the initial cluster model and threading templates, as well as the intrinsic knowledge-based potentials, were employed to direct a subsequent round of structure assembly simulations. The lowest energy structure was selected, which was then subjected to full-atom refinement. Since its first emergence in the CASP7, I-TASSER has consistently achieved top rankings among automated protein structure prediction servers in subsequent CASP experiments [66]. Another example is RosettaCM [94], that assembled structures using integrated torsion space-based and Cartesian space template fragment recombination, loop closure by iterative fragment assembly and Cartesian space minimization, and high-resolution refinement.

2.2. Fragment Assembly Simulation Methods for Free Modeling (FM)

Theoretically, all-atom molecular dynamics (MD) simulations are able to predict protein structures if the computer is powerful enough. However, modern MD simulations can only deal with proteins of less than ~100 amino acids in size. Thus, 90% of the natural proteins cannot be predicted because of the required computational complexity [95]. Hence, an alternative method, namely free modeling (FM), was proposed to model protein structures. Compared to MD simulations, FM methods employ the coarse-grained protein elements and physics- or knowledge-based energy functions, together with extensive sampling procedures, to construct protein structure models from scratch. In contrast to TBM methods, they do not depend on global templates. Hence, they are commonly referred to as ab initio or de novo modeling approaches [17,19]. Since the nature of coarse-grained protein leads to inherent inaccuracies, FM methods, historically, have not achieved levels of accuracy comparable to those of TBM methods, if the global templates are available.

State-of-the-art FM methods have evolved to assemble protein fragments [96]. These fragment assembly techniques assume that protein fragments extracted from the PDB covered most of the conformation of protein folding. Thus, the sampling space was sharply narrowed down. Their implementation involves generating a set of fixed-length (9 residues) and variable-length (15–25 residues) fragments from a repository of known 3D structures (as shown in Figure 3). These fragments are subsequently linked, rotated, and scored to find the global minimum state. This methodology of fragment assembly serves to reduce the exploration of conformational space while ensuring the coherent formation of local structures within the assembled fragments.

The first version of Rosetta modeling software, released in 1997, is one of the most well-known FM methods developed by David Baker's group [17]. Rosetta utilized a three- and nine-residue fragment database for assembly. Particularly, the fragments were selected by quantifying the profile–profile and secondary structure similarity between the query sequence and fragment database within a defined window size. The fragments were simplified to backbone atoms and side-chain centers, and subsequently conducted by simulated annealing Monte Carlo simulations, which exchanged the backbone torsion angles with those of one of the highly scored fragments in the database. A centroid energy function was utilized to guide the simulation, incorporating various factors, such as helix-strand packing, strand pairing, solvation, van der Waals interactions, the radius of gyration, the arrangement of strands into sheets, and interactions between residue pairs. Conformations that exhibited favorable local interactions and possessed protein-like global properties during the simulation were clustered based on their structural similarity, and the final structure was obtained from the center of the largest cluster.

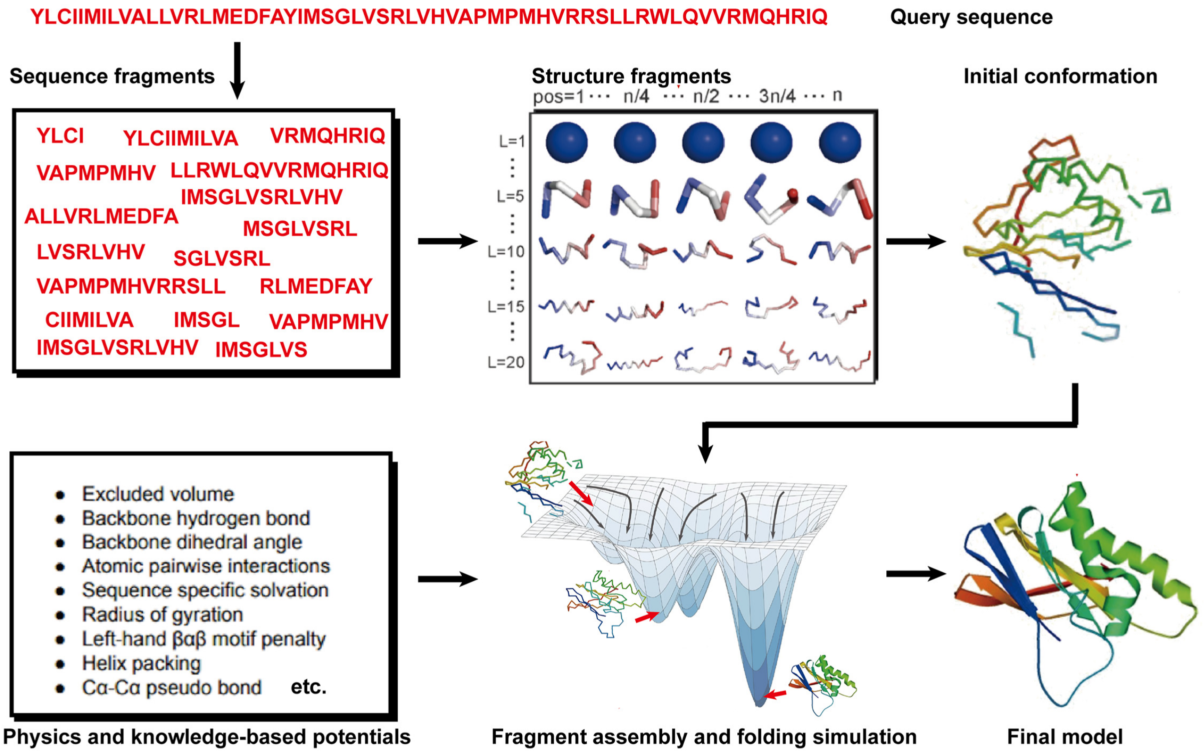

Figure 3. Illustration of free modeling (FM) methods. Starting from a query sequence, local fragments are identified from databases of solved protein structures, using profile-based threading methods. These fragments are subsequently utilized to construct full-length structural models, guided by physics- or knowledge-based energy potentials.

QUARK is another state-of-the-art FM method developed by Yang Zhang's group [19]. Unlike the conventional fragment assembly methods, QUARK utilized distinct methodologies for fragment generation and energy function design. It integrated a distance-based profile energy term, estimating and restricting the distance between two residues by considering inter-residue distances from fragments sourced from the same PDB structures. Additionally, QUARK incorporated 11 diverse conformational movements, improving the efficiency of the conformational sampling procedure, alongside the fragment replacement movement. Today, both the QUARK and Rosetta methods have achieved levels of accuracy comparable to those of TBM methods, and are particularly useful when the protein templates are not available.

2.3. Contact-Based Protein Structure Prediction

A contact map for a protein of length L is defined as a symmetric, binary $L \times L$ matrix. Each element in the matrix represents a binary value, signifying whether the residues form a contact (Cβ-Cβ distance (Cα for glycine) < 8 Å) or not. Since the concept of contact was first brought up, many attempts were made to predict contacts based on correlated mutations in MSAs [97–99]. The hypothesis behind these approaches was that residue pairs that are in contact in 3D space would exhibit correlated mutation patterns, also known as co-evolution (Figure S1), because there is evolutionary pressure to conserve the structures of proteins. A widely used type among these methods is the direct coupling analysis (DCA) method, which considers the full set of pairwise interactions instead of evaluating residues individually. This approach has obtained improved performance compared to mutual information-based methods [99].

In the early 2010s, an increasing number of predictors began integrating deep learning architectures into their prediction methods. A breakthrough occurred in 2017, when Xu's group introduced RaptorX-Contact [22], which revolutionized contact prediction by integrating deep residual convolutional neural networks (ResNets [83]). A Residual Neural Network incorporates an identity map of the input to the output of the convolutional layer, facilitating smoother gradient flow from deeper to shallower layers and enabling training of deep networks with numerous layers. RaptorX-Contact's utilization of deep ResNets, featuring approximately 60 hidden layers, led to a significant performance leap, outstripping other methods [66]. The introduction of deep ResNets, consisting of approximately 60 hidden layers, enabled RaptorX-Contact to significantly outperform other methods [66]. Following RaptorX-Contact's paradigm, several similar methods, like TripletRes [100,101], have emerged.

Due to the latest advances in residue–residue contact prediction, contact-guided protein structure prediction methods have been developed and are becoming more and more successful. The idea of contact-based protein structure prediction methods is described in Figure 4. Starting from a query sequence, an MSA is first generated by searching through databases. The MSA is then used as the input for deep learning methods to predict a contact map. Finally, the contact potential derived from the predicted contact map is used in a folding simulation to predict the final model.

An example of contact-based protein structure prediction methods is CONFOLD2 [102], which builds models using various subsets of input contacts to explore the fold space under the guidance of a soft square energy function, and then clusters the models to obtain the top five models.

The efficacy of deep learning-based contact map prediction was clearly demonstrated by C-I-TASSER and C-QUARK during CASP13, where they ranked in the top two positions among automated servers [23]. These two servers, extended from the classic I-TASSER and QUARK frameworks, incorporated contact maps derived from TripletRes [100,101], ResPRE [103], and various deep learning-based predictors into their simulations. The inclusion of these deep learning restraints significantly enhanced modeling accuracy, particularly for targets lacking easily identifiable template structures [23].

2.4. Distance-Based Protein Structure Prediction

From the definition of contact map prediction, a more detailed extension is distance map prediction. The distinction lies in contact map prediction entailing binary classification, whereas distance map prediction generally estimates the likelihood of the distance between residues falling within various bins (despite attempts made to directly predict real-value distances [104]). Distance map prediction gained significant prominence in the field during CASP13 in 2018, when RaptorX-Contact [22], DMPfold [105], and AlphaFold [106] extended the application of deep ResNets from contact prediction to distance prediction. Among these predictors, AlphaFold, created by Google DeepMind, exhibited superior performance in tertiary structure modeling, as it was ranked as the top one among all groups in CASP13. Leveraging co-evolutionary coupling information extracted from an MSA, AlphaFold employed a deep residual neural network, comprising 220 residual blocks, to predict the distance map for a target sequence, which was subsequently used to assemble protein models. Figure 5 shows the basic steps of distance-based protein structure prediction methods.

A further expansion beyond distance prediction is the prediction of inter-residue torsion angle orientations. The significance of orientation-dependent energy functions serves a dual purpose: biologically, certain residue–residue interactions necessitate not only proximity in distance but also specific orientations between the residue pairs, such as beta strand pairing. From a mathematical standpoint, the inclusion of torsion angle information is crucial, as distance data alone cannot distinctly discern between a pair of mirrored structures, rendering it impossible to uniquely determine the geometry of a structure.

Due to the significance of inter-residue orientations, numerous structure prediction methodologies have integrated them into their workflows. For instance, trRosetta [25,107] has included orientation information by employing a deep residual neural network to predict both pairwise residue distances and inter-residue orientations, based on co-evolutionary information. In CASP14, several leading groups, including D-I-TASSER [108] and D-QUARK [108], incorporated orientation and distance restraints predicted by deep residual neural networks. Moreover, the top CASP14 server group, D-I-TASSER, utilized DeepPotential's residual neural network to predict hydrogen bond networks and integrated these hydrogen bonding restraints into its structural assembly simulations. Notably, the deep learning-based hydrogen bond network prediction significantly enhanced modeling accuracy for CASP14 targets, particularly those lacking homologous templates [108].

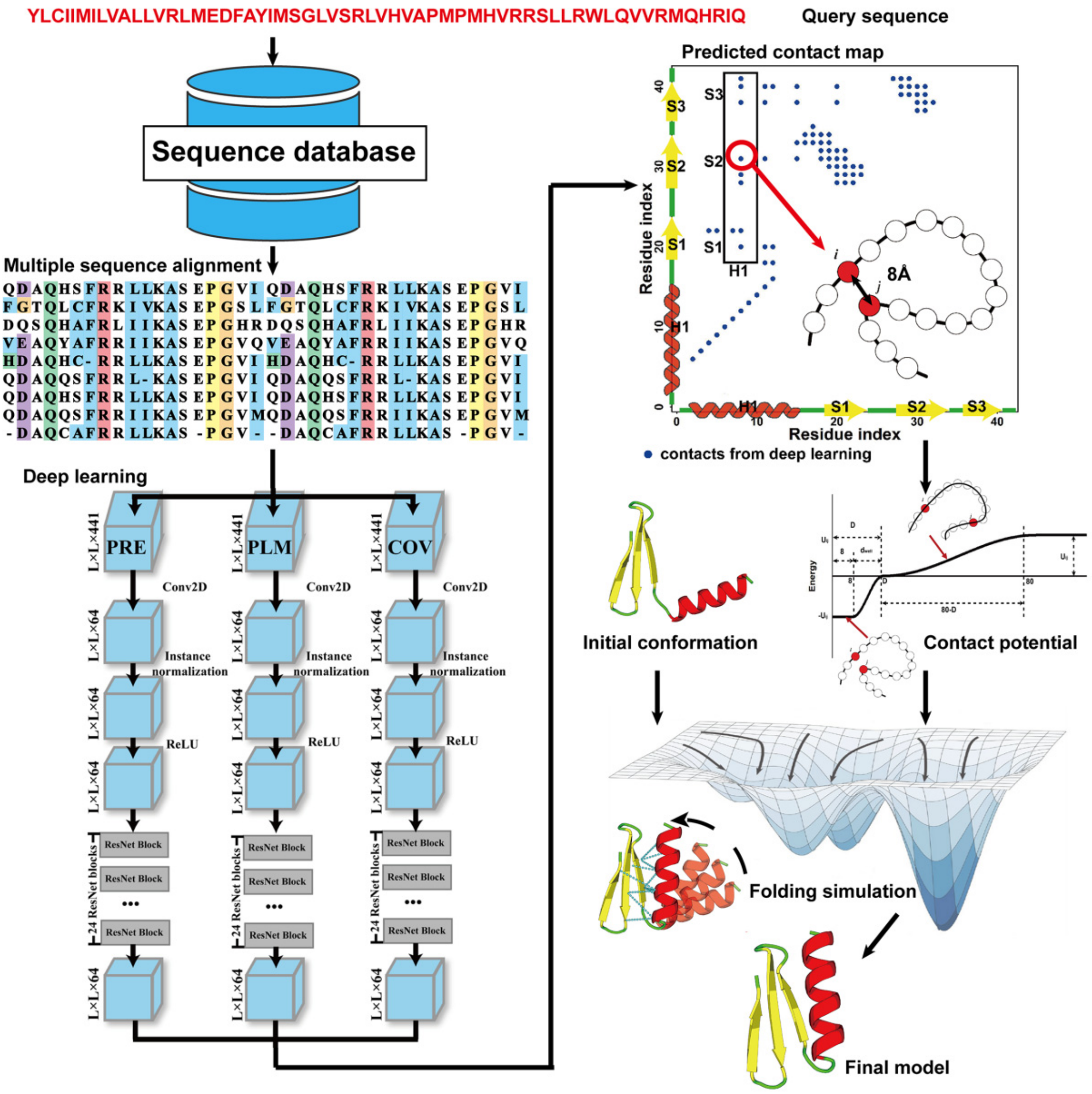
YLCIIMILVALLVRLMEDFAYIMSGLVSRLVHVAPMPMHVRRSLLRWLQVVRMQHRIQ
Query sequence
Sequence database
Multiple sequence alignment
QDAQHSFRRLLKASEPGVIQDAQHSFRRLLKASEPGVI
FGTQLCFRKIVKASEPGSLFGTQLCFRKIVKASEPGSL
DQSQHAFRLIIKASEPGHRDQSQHAFRLIIKASEPGHR
VEAQYAFRRIIKASEPGVQVEAQYAFRRIIKASEPGVQ
HDAQHC-RRLLKASEPGVIHDAQHC-RRLLKASEPGVI
QDAQQSFRRL-KASEPGVIQDAQQSFRRL-KASEPGVI
QDAQHSFRRLLKASEPGVIQDAQHSFRRLLKASEPGVI
QDAQQSFRRIIKASEPGVMQDAQQSFRRIIKASEPGVM
-DAQCAFRRLLKAS-PGV--DAQCAFRRLLKAS-PGV-
Deep learning
L×L×441
PRE
PLM
COV
Conv2D
L×L×64
Instance normalization
ReLU
24 ResNet blocks
ResNet Block
•••
Predicted contact map
S1
S2
S3
H1
Residue index
0
10
20
30
40
8Å
i
j
● contacts from deep learning
Initial conformation
Energy
D
8
d_{well}
U_{ij}
$-U_{ij}$
80
80-D
Contact potential
Folding simulation
Final model

Figure 4. Illustration of contact-based protein structure prediction methods. Starting from a query sequence, an MSA is first generated by searching through databases. The MSA is then used as the input of deep learning methods to predict a contact map. Finally, the contact potential derived from the predicted contact map is used in a folding simulation to predict the final model.

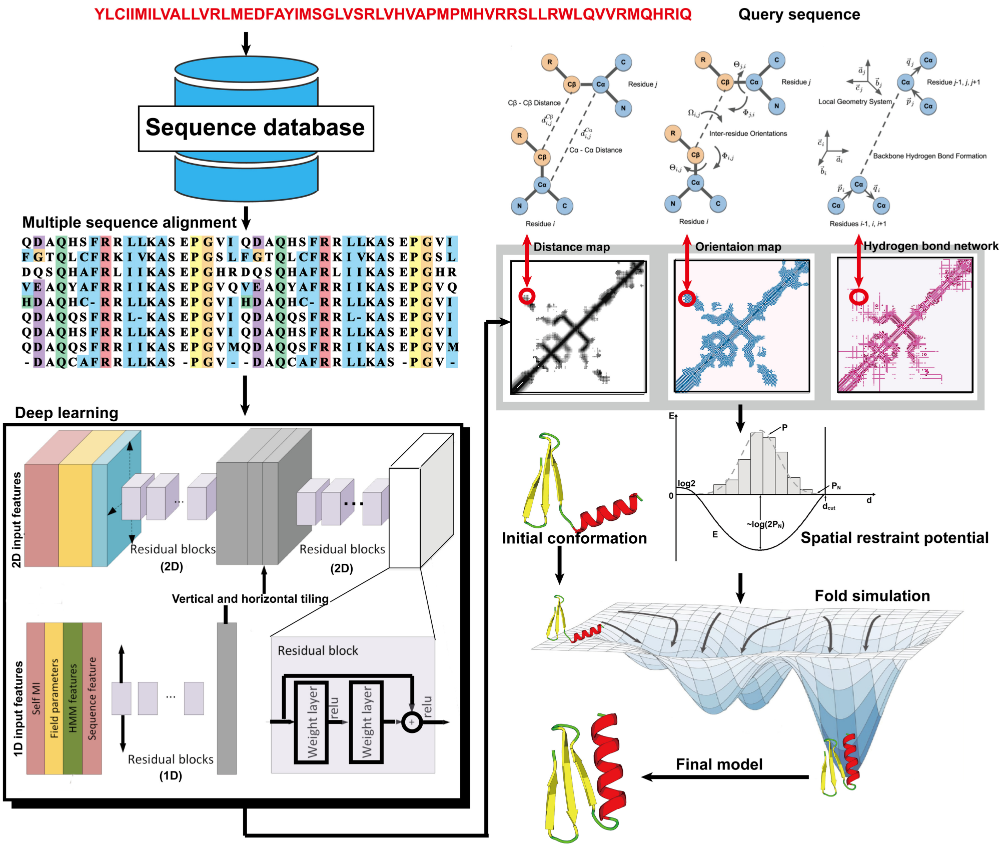

Figure 5. Illustration of distance-based protein structure prediction methods. Starting from a query sequence, an MSA is first generated by searching through databases. Then, the MSA is fed into deep neural networks to predict spatial restraints, such as distance maps, inter-residue orientations, and hydrogen bond networks. Finally, the final structural model is constructed by employing the potentials extracted from the predicted spatial restraints in a folding simulation to identify the lowest energy structure.

2.5. End-to-End Protein Structure Prediction

AlphaFold2 achieved remarkable modeling accuracy and substantially addressed the challenge of predicting the structures of single-domain proteins in CASP14 [109]. The success of AlphaFold2 can be attributed, in part, to its unique "end-to-end" learning approach. This end-to-end learning approach eliminates the need for complex folding simulations, allowing deep neural networks, such as 3D equivariant transformers in AlphaFold2, to predict structural models directly.

AlphaFold2 adopted a novel architecture that is quite different from those of previous methods, including the first version of AlphaFold, to accomplish end-to-end structure prediction. The architecture of AlphaFold2 includes the following two primary components: the Trunk Module, which utilizes self-attention transformers to process input data consisting of the query sequence, templates, and MSA; and the Structure (or Head) Module, which employs 3D rigid body frames to directly generate 3D structures from the training components [110].

Despite its breakthrough in accuracy and performance, AlphaFold2 has notable limitations, such as increased time consumption with longer protein lengths. To address these challenges, several faster artificial intelligence-driven protein folding tools, based on AlphaFold2, have been developed [111–113]. For example, ColabFold [111] improved the speed of protein structure prediction by integrating MMseqs2′s efficient homology search (Many-against-Many sequence searching) [114] with AlphaFold2 [110]. OpenFold [112], a trainable and open-source implementation of AlphaFold2 using PyTorch [115], achieved enhanced computational efficiency with reduced memory usage, thereby facilitating the prediction of exceedingly long proteins on a single GPU. Similarly, Uni-Fold [113] redeveloped AlphaFold2 within the PyTorch framework and reproduced its original training process on a larger set of training data, achieving comparable or superior accuracy and faster speed. Collectively, these developments represent significant strides in enabling rapid and accurate predictions of protein structures.

Tables 1 and 2 show both the domain-level and full-length-level comparisons of TM-scores among AlphaFold2 and its follow-up methods on CASP14 targets (target details are shown in Table S8). The domain-level targets (or domains) are further classified as "TBM-easy", "TBM-hard", "FM/TBM", or "FM" by CASP, depending on the availability and quality of PDB templates for each domain, wherein "TBM-easy" domains have readily identifiable, high-quality templates and "FM" domains typically lack homologous templates in the PDB. To simplify the analysis, "TBM-easy" and "TBM-hard" domains have been merged into "TBM" domains, and "FM/TBM" and "FM" domains into "FM" domains. Here, TM score is a sequence length-independent metric that ranges from [0, 1], in which a score >0.5 indicates that the predicted and native structures share the same global topology [116,117]. From the tables, AlphaFold2 showed excellent performance, only slightly worse than Uni-Fold and ColabFold, especially on FM targets, because of the larger number of training data (that may include CASP14 targets) used in Uni-Fold and the improved MMseqs2-based MSA construction used in ColabFold. Furthermore, AlphaFold2 had an average TM-score of 0.8871 on domain-level assessments (Table 1), but only 0.8514 on full-length-level assessments (Table 2). This is because the full-length-level assessments account for multi-domain targets, whereas AlphaFold2 still needs to be improved. Similar trends can be seen for other AlphaFold2-based methods, indicating that AlphaFold2 and its follow-up methods still need to improve their multi-domain protein structure predictions, even though they have excellent performance on single-domain proteins.

In addition to AlphaFold2 and its related methods, Baker's group has developed RoseTTAFold [118], which used a three-track network to process sequence, distance, and coordinate information simultaneously, and achieved high prediction accuracy at CASP14, ranking only behind AlphaFold2.

Table 1. Comparison of domain-level modeling results by AlphaFold2-based methods and protein language model (PLM)-based methods for different domain types on the 91 CASP14 domains. The original CASP "TBM-easy" and "TBM-hard" domains are categorized as "TBM" domains, while the "FM/TBM" and "FM" domains are categorized as "FM" domains in this analysis. Here, AlphaFold2-Single is the default AlphaFold2 pipeline, with the only query sequence as the input MSA. *p*-values were calculated between TM-scores by AlphaFold2 and others using paired one-sided Student's *t*-tests. #{TM > 0.5} is the number of targets with a TM-score > 0.5.

Method	Method Type	Domain Type	TM-Score	*p*-Value	#{TM > 0.5}
AlphaFold2	AlphaFold2-based	All	0.8871	-	88
		TBM	0.9325	-	54
		FM	0.8207	-	34
ColabFold		All	0.8846	3.29×10^{-2}	88
		TBM	0.9255	6.65×10^{-3}	54
		FM	0.8250	4.25×10^{-1}	34
OpenFold		All	0.8692	3.17×10^{-2}	85
		TBM	0.9199	1.57×10^{-1}	53
		FM	0.7952	5.24×10^{-2}	32
Uni-Fold		All	0.8930	9.93×10^{-1}	88
		TBM	0.9387	9.76×10^{-1}	54
		FM	0.8262	9.26×10^{-1}	34

Table 1. *Cont.*

Method	Method Type	Domain Type	TM-Score	*p*-Value	#{TM > 0.5}
AlphaFold2-Single	PLM-based	All	0.5165	4.06×10^{-16}	40
		TBM	0.6609	5.15×10^{-10}	37
		FM	0.3057	2.18×10^{-11}	3
ESMFold		All	0.7206	4.64×10^{-14}	66
		TBM	0.8481	1.89×10^{-7}	50
		FM	0.5346	1.02×10^{-10}	16
OmegaFold		All	0.6920	4.60×10^{-9}	64
		TBM	0.7944	5.42×10^{-6}	46
		FM	0.5426	2.18×10^{-5}	18

Table 2. Comparison of full-length-level modeling results by AlphaFold2-based methods and protein language model (PLM)-based methods on the 65 CASP14 full-length targets. Here, AlphaFold2-Single is the default AlphaFold2 pipeline, with the only query sequence as the input MSA. *p*-values were calculated between TM-scores by AlphaFold2 and others using paired one-sided Student's *t*-tests. #{TM > 0.5} is the number of targets with a TM-score > 0.5.

Method	Method Type	TM-Score	*p*-Value	#{TM > 0.5}
AlphaFold2	AlphaFold2-based	0.8514	-	60
ColabFold		0.8461	2.37×10^{-1}	61
OpenFold		0.8375	1.46×10^{-1}	59
Uni-Fold		0.8561	9.82×10^{-1}	61
AlphaFold2-Single	PLM-based	0.5164	6.88×10^{-12}	30
ESMFold		0.6676	1.80×10^{-10}	41
OmegaFold		0.6728	4.42×10^{-6}	43

2.6. Protein Language Model-Based Protein Structure Prediction

AlphaFold2 has facilitated the rise of structure prediction performance to new heights, nearly comparable to the accuracy of experimental determination methods since CASP14. Standard protein structure prediction pipelines heavily rely on co-evolution information from MSAs. However, the excessive dependence on MSAs often acts as a bottleneck in various protein-related problems. While model inference in the structure prediction pipeline typically takes a few seconds, the MSA construction step is time-intensive, consuming tens of minutes per protein. This time-consuming process significantly hampers tasks requiring high-throughput requests, like protein design [119]. Therefore, developing an accurate and efficient MSA-free protein structure prediction method holds promise in advancing protein studies.

A large-scale protein language model (PLM) presents an alternative avenue to MSAs for acquiring co-evolutionary knowledge, facilitating MSA-free predictions. In contrast to MSA-based methods, wherein information retrieval techniques explicitly capture co-evolutionary details from protein sequence databases, PLM-based methods embed co-evolutionary information into the large-scale model parameters during training, and allow for implicit retrieval through model inference, wherein the PLM is viewed as a repository of protein information. Furthermore, MSA-based approaches have lower efficiency in information retrieval, relying on manually designed retrieval schemes. Conversely, a PLM-based method showcases heightened efficiency in information retrieval, with retrieval quality predominantly influenced by the model's capacity or parameter size. A lot of pre-trained PLMs have been developed and released for various downstream analyses [85,120], such as SaProt [120], which is a large-scale general-purpose PLM trained on an extensive dataset comprising approximately 40 million protein sequences and structures, and ESM-

2 [85], which was trained on protein sequences from the UniRef database, with up to 15 billion parameters.

Inspired by the progress of PLMs and AlphaFold2, many protein structure prediction methods have been proposed. For example, ESMFold [85], developed by Meta AI, used the information and representations learned by a PLM called ESM-2 to perform end-to-end 3D structure prediction using only a single sequence as input. ESMFold demonstrated comparable accuracy to AlphaFold2 and RoseTTAFold for sequences exhibiting low perplexity and thorough comprehension by PLM. Notably, ESMFold's inference speed was ten times faster than that of AlphaFold2, thereby facilitating efficient exploration of the structural landscape of proteins within practical time frames. OmegaFold [121] predicted the high-resolution protein structure from a single primary sequence alone, using a combination of a PLM and a geometry-inspired transformer model, trained on protein structures. OmegaFold requires only a single amino acid sequence for protein structure prediction and does not rely on MSAs or known structures as templates. Similar to ESMFold, OmegaFold can also scale roughly ten times faster than MSA-based methods, such as AlphaFold2 and RoseTTAFold. HelixFold-Single [119] was an end-to-end MSA-free protein structure prediction pipeline that combined a large-scale PLM with the superior geometric learning capability of AlphaFold2. HelixFold-Single first pre-trained a large-scale PLM with thousands of millions of primary structures, utilizing the self-supervised learning paradigm, and then obtained an end-to-end differentiable model to predict 3D structures by combining the pre-trained PLM and the essential components of AlphaFold2. EMBER3D [122] predicted 3D structure directly from single sequences by computing both 2D (distance maps) and 3D structure (backbone coordinates) from sequences alone, based on embeddings from the pre-trained PLM called ProtT5. EMBER3D exhibited a speed that was orders of magnitude faster than its counterparts, enabling the prediction of average-length structures in mere milliseconds, even on consumer-grade machines.

The benchmark results in Tables 1 and 2 indicate that PLM-based protein structure prediction methods are generally worse than MSA-based methods, although PLM-based methods run very fast. Due to the large scalability of PLM-based methods, they have broad application prospects, and still require further improvements in terms of accuracy.

2.7. Multi-Domain Protein Structure Prediction

Since the advent of AlphaFold2 in the recent CASP14, great progress has been made in protein structure prediction. However, AlphaFold2 and most of the subsequent state-of-the-art methods have mainly focused on the modeling of single-domain proteins, which are the minimum folding units of proteins that fold and function independently. Nonetheless, it is worth noting that several of the CASP14 targets, especially large multi-domain targets, were not predicted with high accuracy, suggesting that further improvements are needed for multi-domain prediction [123]. As shown in Tables 1 and 2, AlphaFold2 had an average TM-score of 0.8871 on domain-level assessments, but only 0.8514 when considering multi-domain targets. This is because the full-length-level assessments account for multi-domain targets, where AlphaFold2 still needs to be improved. In fact, more than two-thirds of prokaryotic proteins and four-fifths of eukaryotic proteins contain two or more domains [124]. Therefore, determining the full-length structures of multi-domain proteins is highly required.

A common approach to multi-domain protein structure modeling is to split the query sequence into domains and generate models for each individual domain separately. The individual domain models are subsequently assembled into full-length models, usually under the guidance of other homologous multi-domain proteins from the PDB. Such domain assembling methods can be divided into the following two categories: linker-based domain assembly and inter-domain rigid body docking. Linker-based methods, such as Rosetta [125] and AIDA [126], primarily focus on the construction of linker models by exploring the conformational space, with domain orientations loosely constrained by physical potential from generic hydrophobic interactions. Docking-based methods, such as DEMO [127,128] and SADA [129], assemble the single domain structure via rigid body docking, which is essentially a template-based method that guides domain assembly by detecting available templates.

Furthermore, some fully automated pipelines [130] for multi-domain protein structure prediction from sequences alone have been developed based on this idea. For example, I-TASSER-MTD first predicted domain boundaries from sequences by FUpred [131] and ThreaDom [132]. Then, single-domain structural models were folded by the original version of D-I-TASSER [108] guided by deep-learning spatial restraints [100,101]. Finally, DEMO [127,128] was used to perform multi-domain structure assembly.

Note that the performance of common protein structure prediction methods relies, to some extent, on the quality of the MSA or the homologous template [66]. However, homologs available in the PDB may be fewer for multi-domain proteins, which may further affect the performance of multi-domain protein structure prediction. Thus, some threading-based methods, such as LOMETS3 [77], have been developed to increase template recognition and alignment accuracy for multi-domain proteins. LOMETS3 performed three steps of domain boundary prediction, domain-level template identification, and full-length template/model assembly, which can help better detect distant homologous templates for multi-domain proteins [77]. Furthermore, the DeepMSA2 [133] algorithm has been proposed to generate deeper MSAs, facilitating the improvement in MSA quality for multi-domain protein structure prediction.

Aside from the challenges presented by shallow MSAs, another significant limitation in multi-domain protein structure prediction is accurately modeling the orientation between different domains. Some efforts have been made to improve the inter-domain orientation problem in multi-domain protein structure prediction. For example, DeepAssembly [134] used a population-based evolutionary algorithm to assemble multi-domain proteins, leveraging inter-domain interactions inferred from a developed deep learning network. E2EDA [135] was an end-to-end domain assembly method based on deep learning. It first predicted inter-domain rigid motion using an attention-based deep learning model. Subsequently, these predicted rigid motions were translated into inter-domain spatial transformations to allow for the direct assembly of full-chain models. The final stage involved selecting the best model from multiple assembled models, guided by a specific scoring strategy.

Furthermore, the latest version of the D-I-TASSER pipeline has been developed by integrating all aforementioned strategies to improve multi-domain protein structure predictions. D-I-TASSER first generated MSAs by DeepMSA2 [133], which were then used for template identification by LOMETS3 [77] and spatial restraint prediction by AlphaFold2, AttentionPotential [133], and DeepPotential [136], on both the full-length level and the domain level, with the aid of a multi-domain handling module that incorporated FUpred [131], ThreaDom [132], and DEMO2 [127]. Unlike I-TASSER-MTD, which attempted to assemble domain-level models into the full-length model, D-I-TASSER directly predicted the full-length atomic model from both full-length-level inputs and domain-level assembled inputs, that is, the templates and spatial restraints, through the Replica Exchange Monte Carlo (REMC) folding system [91–93]. In this way, the inter-domain orientation information contained in full-length-level inputs can be used to construct the final model. D-I-TASSER (named as "UB-TBM") participated in the CASP15 "Inter-domain Modeling" Section, which corresponds to multi-domain structure prediction. D-I-TASSER outperformed all other groups in terms of the Z-score sum, calculated by the CASP Assessors (Figure 6). In particular, the Z-score sum of D-I-TASSER (35.53) was 42.3% higher than that of the second-best performing group (24.96) (see https://predictioncenter.org/casp15/zscores_interdomain.cgi, accessed on 10 December 2023).

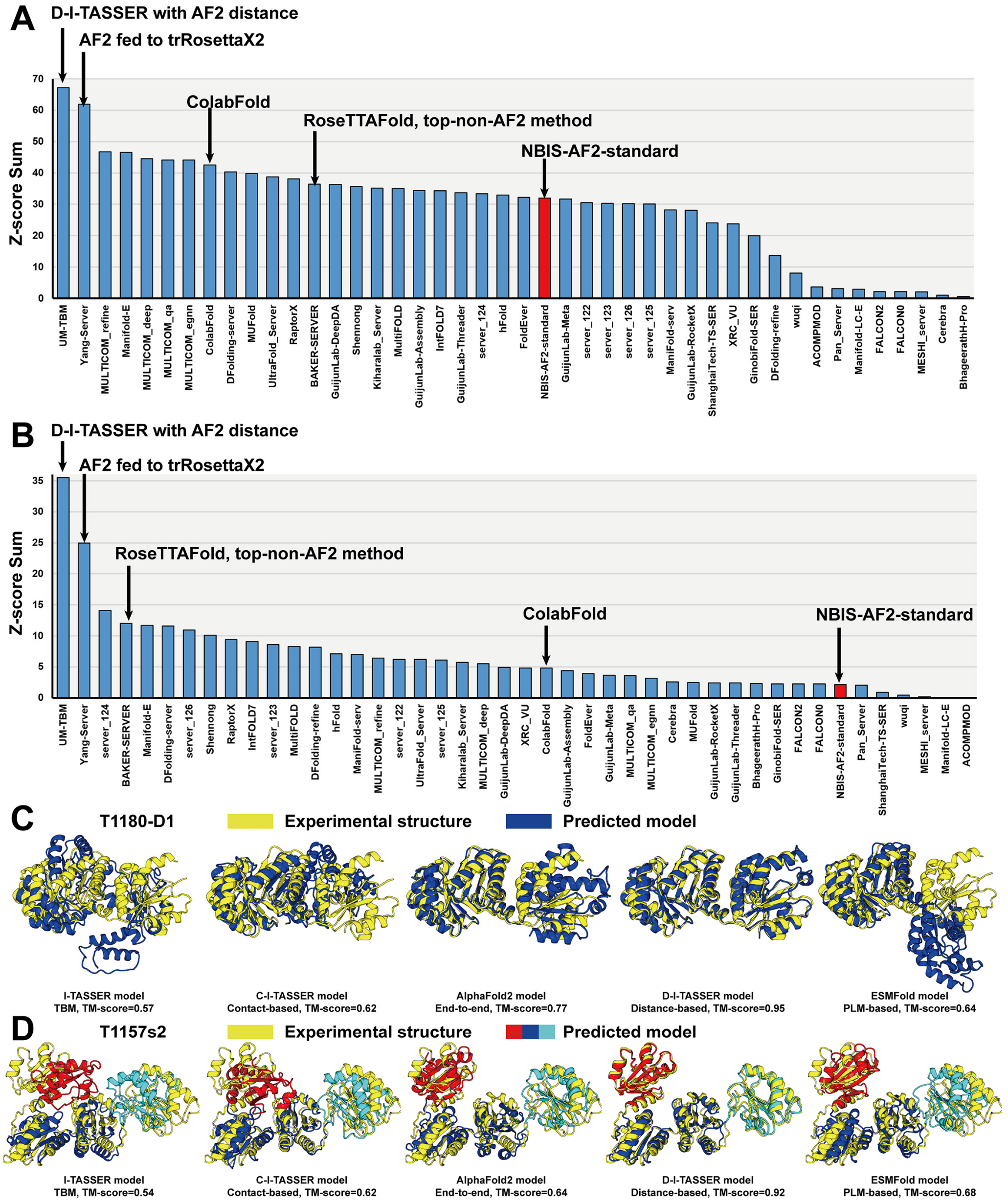

Figure 6. Protein structure prediction results in CASP15. (**A**,**B**) Sums of Z-scores for the top 44 registered server groups in the (**A**) "Regular Modeling" and (**B**) "Inter-domain Modeling" Sections in CASP15. The public version 2.2.0 of the AlphaFold2 server (registered as "NBIS-AF2-standard") is marked in red. (**C**,**D**) The modeling performance of I-TASSER (a template-based modeling (TBM) method), C-I-TASSER (a contact-based method), D-I-TASSER (a distance-based method), AlphaFold2 (an end-to-end method), and ESMFold (a protein language model (PLM)-based method) on representative examples of (**C**) CASP15 single-domain target T1180-D1 and (**D**) CASP15 multi-domain target T1157s2. The single-domain predicted models are depicted in blue, the multi-domain predicted models are marked by red, blue, and cyan to distinguish different domains, and the superposed experimental structures are represented by yellow.

2.8. CASP and Most Recent CASP Results

The Critical Assessment of Protein Structure Prediction (CASP) was established in 1994, by Professor John Moult and others from the University of Maryland, and has taken place every other year since then [137]. Its purpose is to provide an objective evaluation of protein structure prediction technologies within the field of protein structure prediction. Employing a rigorous double-blind prediction mechanism, it is viewed as the gold standard for assessing protein structure prediction techniques and is regarded in the industry as the "Olympics of protein structure prediction".

In order to fairly evaluate protein structure prediction methods, CASP assessors have incorporated and designed multiple measures. Two widely used evaluation measures by CASP are the TM-score and the global distance test score (GDT score). The TM-score between the model and the experimental structure is usually used to assess the global quality of a structural model [138]. The TM-score ranges between 0 and 1, with TM-scores > 0.5 indicating that the structure models have the same fold defined in SCOP/CATH [117]. The GDT score is calculated by GDT = (GDT_P1 + GDT_P2 + GDT_P4 + GDT_P8)/4, where GDT_Pn indicates the percent of residues under the distance cut-off $\leq$ n Å [139]. The GDT score primarily focuses on assessing the backbone modeling quality of a protein. With the substantial enhancement in prediction accuracy witnessed since the advent of AlphaFold2 in CASP14, more and more measures for assessing side-chain modeling quality have been introduced. For instance, SC_error is a measure used for assessing side-chain modeling quality, while MolProbity is a comprehensive scoring function used for assessing the non-physical area of the model (i.e., atom clash, rotamer outlier, favored Ramachandran, etc.).

According to the rules of CASP, all participating methods are categorized into the following two groups: server-based and human-based. Participants in the server-based group have a limited window of 72 h for structure prediction, while those in the human-based group are allotted 3 weeks, allowing for manual intervention. This signifies that the server-based group relies solely on computer predictions; hence, the competitive difficulty in this category is often higher than in the human-based groups.

Starting from CASP7, the proteins modeled during CASP have been classified as TBM, TBM-easy, TBM-hard, FM/TBM, or FM, depending on the availability and quality of PDB templates for each target, where TBM-easy targets have readily identifiable, high-quality templates, and FM targets typically lack homologous templates in the PDB. For the purpose of analyses, TBM, TBM-easy, and TBM-hard are often regarded as TBM targets, and FM/TBM and FM are treated as FM targets.

Starting from CASP12, protein complex prediction has been included in CASP as an independent assessment category, called the protein assembly category. Protein complex modeling is distinguished from the classical protein–protein docking, where two protein subunits, named the ligand and the receptor, are in contact through a single interface. In the CASP protein assembly assessment, predictions of full-length protein complexes involve predictions of both individual protein–protein interfaces and overall complex topology.

Starting from CASP13, deep learning techniques have achieved significant breakthroughs, markedly enhancing the accuracy of protein tertiary structure prediction.

In CASP13, the adoption of distance map prediction began to play a pivotal role in guiding protein structure prediction. Notable examples include RaptorX-Contact [22], DMPfold [105], and AlphaFold [106], which employed deep Residual Networks (ResNets) from contact prediction to distance prediction, significantly boosting predictive modeling performance. In particular, AlphaFold, developed by Google DeepMind, was ranked as the top method in tertiary structure modeling among all groups in CASP13. However, the majority of other groups continued to rely on contact prediction information for guiding protein structure prediction. Due to the remarkable accuracy of deep learning-based contact map predictions, even contact-based protein structure prediction methods also achieved excellent performance. For instance, C-I-TASSER and C-QUARK were ranked as the top two automated servers during CASP13 [23].

The effectiveness of distance prediction, as demonstrated in CASP13, has led to its widespread applications in various structure prediction methodologies. A promising example is trRosetta [25,107], which employed a deep residual neural network to predict both pairwise residue distances and inter-residue orientations for guiding protein structure prediction. Following the inspiration from trRosetta, numerous groups in CASP14 incorporated orientation and distance constraints predicted by deep residual neural networks into their protein structure prediction processes. Among these methods, D-I-TASSER [108] and

D-QUARK [108] were two top CASP14 servers from Yang Zhang's group. D-I-TASSER, in particular, leveraged deep learning-based hydrogen bond network prediction to guide protein structure prediction, significantly improving modeling accuracy for CASP14 targets, especially those lacking homologous templates [108]. More importantly, AlphaFold2 represented a groundbreaking shift by employing an end-to-end deep learning approach to protein structure prediction, and facilitated the rise of predictive performance to unprecedented levels, regularly competitive with experimental structures in CASP14.

In CASP15, following the release of the AlphaFold2 codes, most groups adopted the AlphaFold2 framework for their structure predictions, resulting in outstanding performance across the board. Figure 6A,B list the sums of Z-scores, calculated by the CASP Assessors, for the top 44 CASP15 server groups that participated in the CASP15 "Regular Modeling" (https://predictioncenter.org/casp15/zscores_final.cgi?formula=assessors&gr_type=server_only, accessed on 10 December 2023) and "Inter-domain Modeling" (https://predictioncenter.org/casp15/zscores_interdomain.cgi, accessed on 10 December 2023) Sections, which correspond to single- and multi-domain structures, respectively. Here, we only show the results from server groups because the human group results may incorporate experience and expertise, which may be unfair for evaluating different protein structure prediction methods. In CASP15, due to the release of the AlphaFold2 standalone package, most of the participant methods were AlphaFold2-based methods. In particular, the top five performing methods were all based on AlphaFold2, with their own modifications, such as incorporating AlphaFold2 with other simulation pipelines, using diverse MSAs, and fine-tuning AF2 refinements; thus, they acquired much better performance than the default AlphaFold2 (registered as the "NBIS-AF2-standard" group). The top non-AlphaFold2 method was based on RoseTTAFold2 (registered as the "BAKER" group), which had good predictive performance on multi-domain proteins. In Figure 6C,D, we used representative examples of a single-domain target, T1180-D1, and a multi-domain target, T1157s2, from CASP15 to highlight the modeling performance of different types of methods, including a template-based modeling (TBM) method, I-TASSER, a contact-based method, C-I-TASSER, a distance-based method, D-I-TASSER, an end-to-end method, AlphaFold2, and a protein language model (PLM)-based method, ESMFold. The TBM method exhibited the worst performance, with TM-scores of 0.57 and 0.54 for the single-domain and the multi-domain targets, respectively. The contact-based method also showed limited accuracy for both targets. AlphaFold2, the recently developed end-to-end method, demonstrated improved performance on the single-domain target (TM-score = 0.77) but slightly reduced efficacy on the multi-domain target (TM-score = 0.64), highlighting the inherent challenges in multi-domain protein structure prediction. Notably, the latest version of D-I-TASSER achieved remarkable predictive accuracy for both single-domain and multi-domain targets by carefully integrating the AlphaFold2 pipeline with a multi-domain handling module. On the other hand, despite its rapid execution, the PLM-based method exhibited suboptimal performance, particularly on the single-domain target.

In particular, CASP15 introduced a new category, ligand prediction, where participants were provided with both protein (or RNA) and ligand data to generate 3D structural models for the corresponding protein/RNA–ligand complexes [140]. All leading groups in this category adopted similar methodologies, which started from a search in the PDB for similar ligands and binding pockets. Following this, the identified PDB binding pockets were superimposed onto the AlphaFold2 structures of the target proteins. This superposition facilitated the generation of an initial pose for the ligand. To further refine and evaluate these alignments, various conventional methods and machine learning techniques were employed.

For example, the CoDock approach [141] combined template-based modeling with a convolutional neural network (CNN)-based scoring function to predict ligand binding. The Zou group [142] adopted a similar strategy, integrating the physicochemical molecular docking method AutoDock Vina [143] with the ligand similarity methodology SHAFTS [144]. In the Alchemy_LIG team [145] protein structures were constructed using

AlphaFold2, and ligands were docked utilizing the AutoDock Vina docking method and a machine learning model trained to detect native binding modes. The ClusPro group [146] employed AlphaFold2 for constructing monomer protein structures and created multimeric assemblies via a template-based docking algorithm, ClusPro LigTBM [146], for general ligand placement, alongside the Glide program [147], for direct docking in cases when no templates were found.

While docking approaches utilizing templates from the PDB demonstrated superior performance, it is important to recognize that the excellent performance of these template-based methods was not uniformly observed across all CASP15 targets [140]. Furthermore, it is noteworthy that state-of-the-art deep learning techniques have yet to be extensively employed in the realm of protein–ligand structure predictions, representing a significant and promising avenue for future research.

2.9. AlphaFold Protein Structure Database (AlphaFold DB)

The AlphaFold Protein Structure Database (AlphaFold DB, https://alphafold.ebi.ac.uk, accessed on 10 December 2023), created in partnership between DeepMind and the EMBL-European Bioinformatics Institute (EMBL-EBI), is a freely accessible database of high-accuracy protein structure predictions by the scientific community [148]. Powered by AlphaFold2 of Google DeepMind, AlphaFold DB provides highly accurate protein structure predictions, competitive with experimental structures. The latest AlphaFold DB release contains over 200 million entries, providing broad coverage of UniProt [149], which is the standard repository of protein sequences and annotations. AlphaFold DB provides individual downloads for the human proteome and for the proteomes of 47 other key organisms important in research and global health. AlphaFold DB also provides a download for the manually curated subset of UniProt. The prediction results of AlphaFold DB can be accessed through several mechanisms, as follows: (i) bulk downloads (up to 23 TB) via FTP; (ii) programmatic access via an application programming interface (API); and (iii) download and interactive visualization of individual predictions on protein-specific web pages keyed on UniProt accessions.

The AlphaFold DB's release of a multitude of novel protein structures has provided bioinformaticians across the globe with a rich repository of data. Developers specializing in protein structure analysis tools are leveraging this influx of accurate models, leading to numerous significant breakthroughs in protein-related fields.

For example, the AlphaFold DB, through its accurate prediction of protein structures, offers a robust foundation for understanding how different ligands might interact with various proteins, which is pivotal in identifying potential drug targets, aiding in the design of novel pharmaceuticals, and contributing to a broader understanding of biological functions. In this context, several methods have been developed. AlphaFill, for instance, was developed to enrich the models in the AlphaFold DB by "transplanting" ligands, co-factors, and ions, based on sequence and structure similarity [150]. Similarly, Wehrspan et al. investigated the binding sites for iron–sulfur (Fe-S) clusters and zinc (Zn) ions within predicted structures in AlphaFold DB [151]. With the utilization of the AlphaFold DB, PrankWeb3 was able to predict protein–ligand binding sites in situations where no experimental structure is available [152].

Another recent application of AlphaFold DB was related to post-translation modifications (PTMs) [153], where structural insights obtained from AlphaFold DB were systematically integrated with proteomics data, particularly large-scale PTM information, aiming to illuminate the functional significance of PTMs.

While the AlphaFold DB has significantly expanded the application and scalability of tools and algorithms for protein-related analyses, effectively analyzing more than a couple of hundred thousand protein structures or models poses a challenge. There is a pressing need to develop novel approaches capable of managing the unanticipated and rapid growth of available models. Notably, state-of-the-art tools such as FoldSeek [154] and 3D-AF-Surfer [155] have already been developed, aiding researchers in searching through extensive repositories of protein structures to identify hits with structural similarity to a provided input structure. Leveraging high-throughput structural similarity searches facilitates classification problems, such as assigning structural CATH domains to AlphaFold models [156].

However, many limitations and challenges still remain for AlphaFold DB, such as predicting multi-domain protein structures, and predicting structures for very large proteins (longer than 5000 residues) [157].

3. Discussion and Perspective

Since Anfisen first demonstrated that the information encoded in a protein sequence determines its structure [1], the prediction of protein structures starting from amino acid sequences has remained a challenging problem in structural biology. A number of methods have been proposed to address the problem of protein structure prediction.

The traditional approaches for solving the protein structure prediction problem involve template-based modeling (TBM) and template-free modeling (FM) methods. The TBM approaches demonstrate high efficacy when homologous templates are easily identifiable. However, their accuracy significantly decreases in cases where only distantly related templates are available for a target (see Table 3). On the other hand, FM methods are generally limited to folding smaller, non-beta proteins because of the computational complexities inherent in their energy functions and conformational sampling techniques.

Table 3. The advantages and limitations of each type of methods.

Method	Advantages	Limitations
Template-based modeling (TBM)	The methods can achieve high accuracy and adeptly reflect evolutionary relationships when reliable templates are identifiable.	The accuracy of TBM significantly decreases when the available templates are only distantly related to the target protein.
Template-free modeling (FM)	The methods are not limited to the availability of templates and, thus, can be applied to any protein.	The statistical and knowledge-based energy potentials used in FM methods may lead to suboptimal performance if they are inaccurate. Also, these energy potentials contain little residue–residue interaction information.
Contact/distance-based methods	The energy potentials derived from deep learning-based restraints (contacts or distances) contain high-quality residue–residue interaction information.	The deep learning-based restraints (contacts or distances) and the final structural models are optimized separately, which may be difficult for improving overall accuracy. Additionally, the requirement for MSA inputs poses a challenge for distance-based methods, especially in cases in which high-quality MSAs are difficult to obtain.
End-to-end methods	The deep learning-based restraints (contacts or distances) and the final structural models are optimized together, resulting in their high accuracy in single-domain proteins.	Such methods have shown limitations in accurately predicting the structures of multi-domain proteins, especially for proteins with few known homologs.
Protein language model (PLM)-based methods	These methods have high scalability and computational efficiency, since they do not rely on MSA inputs. Also, their performance is relatively better for orphan proteins.	PLM-based methods currently suffer from relatively low accuracy in structure prediction.
Multi-domain protein structure prediction methods	These methods are well-designed for multi-domain proteins, with high performance to balance the modeling quality of inter-domain and intra-domain interactions.	These methods face challenges in carefully balancing MSAs for both separate domains and full-length proteins, accurately modeling the orientations between disparate domains, and predicting the accurate domain boundaries.

Recent breakthroughs in deep learning-based restraint prediction and end-to-end folding have significantly revolutionized the field of protein structure prediction. These developments have markedly improved prediction accuracy and the ability to fold proteins that lack corresponding homologous templates in the PDB. In particular, AlphaFold2 and subsequent methodologies have largely tackled the challenge of protein structure prediction at the domain level through the implementation of end-to-end learning and attention-based networks. However, the predictive accuracy of these AlphaFold2-based methods is significantly dependent on the quality of multiple sequence alignments (MSAs). To bypass the over-reliance on MSAs, protein language model (PLM)-based methods have been developed as alternatives to MSAs for acquiring co-evolutionary information, thus enabling MSA-free predictions. Although these PLM-based approaches are notably rapid, due to the absence of MSA construction, their performance still requires further improvements.

It is crucial to note that neither end-to-end methods nor PLM-based methods can predict multi-domain proteins with high accuracy. Consequently, many methods have been designed for multi-domain protein structure predictions in particular. Nevertheless, substantial challenges persist, particularly in the construction of high-quality MSAs and the accurate modeling of orientations between disparate domains. While some advancements have been made to solve these limitations, there remains a need for further improvements in multi-domain protein prediction, as demonstrated by the generally reduced performance in the "Inter-domain Modeling" Section of CASP15 (Figure 6B).

While the majority of structure prediction methods are based on static structures, it is crucial to recognize that proteins often exist in multiple conformational states, intricately linked to their distinctive functional roles. Notably, the understanding of protein conformational states and folding pathways is critically important in drug development. Furthermore, conformational changes are a key concern in protein–ligand prediction. The principal challenge in this area comes from the limited availability of data on protein motion and evolutionary information. With the increasing number of experimental data, it is expected that more and more methods will be developed to address these challenges [158–160]. Particularly, AlphaFold DB, with its remarkable accuracy in predicting protein structures, has facilitated improvements in this field. For instance, AlphaFold2 successfully demonstrated its ability to identify alternative states of known metamorphic proteins with high confidence by clustering a MSA based on sequence similarity, indicating a significant leap forward in understanding protein dynamics [159]. In addition, a recent study introduced a methodology that utilized AlphaFold2 to sample alternative conformations of topologically diverse transporters and G-protein-coupled receptors, which were not included in the AlphaFold2 training dataset [160].

Due to the high accuracy of recent protein structure prediction methods, these methods can effectively help biologists conduct protein structure and function analyses, for example, using protein structure prediction to assist cryo-electron microscopy electron density maps to resolve atomic-level experimental structures [161,162] and analyzing the structural and functional differences of specific proteins from different species through protein structure prediction methods [163]. In particular, during the novel coronavirus pneumonia outbreak at the end of 2019, no protein structures of the virus were initially analyzed. Given the critical role of the viral proteome as a functional carrier, understanding its structure was important for analyzing the mechanism of viral host invasion. Consequently, several research groups have predicted the full proteome of the SARS-CoV-2 virus, as well as the spike protein of the mutant virus [164,165], and made these predictions freely available in databases for biological researchers.

As protein monomer structure predictions have achieved high accuracy, more and more attention has shifted toward protein complex structure predictions and RNA-related structure predictions. For example, advanced protein structure prediction approaches have been extended to protein complex structure prediction [133,166]. Since most proteins cooperate with their protein interaction partners to form a complex for performing their biological functions in biological processes within a living cell, various experimental methods have been proposed to detect protein complexes, such as AlphaFold-Multimer [166] and DMFold-Multimer [133]. A primary challenge in complex prediction lies in the substantial computational resources necessary for the prediction of large, multi-chain proteins. Furthermore, acquiring high-quality MSAs for complexes is also a particularly challenging task.

Another extension of protein structure prediction involves RNA structure prediction [167–169] and RNA–protein complex structure prediction [170], where representative methods include AIchemy_RNA2 [167], DRfold [168], trRosettaRNA [169], and RoseTTAFoldNA [170]. Despite the increasing accumulation of experimental structural data for RNA, the field of RNA or RNA–protein structure prediction is still challenged by the limited availability of RNA sequence and structure databases, as well as the complexities in extracting conservation information from RNA sequences. As demonstrated by CASP 15, deep learning-based RNA structure predictors did not surpass the performance of traditional energy function-based methods because the performance of deep learning-based methods heavily relies on the number of training data available. The accuracy of RNA structure predictions, whether obtained through traditional or deep learning methodologies, is far from satisfactory.

Although AlphaFold2 and many state-of-the-art methods constitute a significant advancement in "solving" the problem of protein structure prediction from sequences, they are not the final answer. There are still challenges met in searching for high-quality MSAs, improving the side-chain modeling quality [171], and so on. Furthermore, challenges in protein complex structure predictions, RNA-related structure predictions, and protein–ligand structure predictions have received growing attention. The rapid progress observed in recent years brings hope that the problems and challenges associated with protein structure prediction could ultimately be solved by leveraging deep learning techniques in the future.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/molecules29040832/s1, Figure S1: An illustration of co-evolutionary information contained in multiple sequence alignments and the corresponding relationships with residue–residue contact prediction; Table S1: Tools for template-based protein structure prediction; Table S2: Tools for template-free (free modeling) protein structure prediction; Table S3: Tools for contact-based protein structure prediction; Table S4: Tools for distance-based protein structure prediction; Table S5: Tools for end-to-end protein structure prediction; Table S6: Tools for protein language model-based protein structure prediction; Table S7: Tools for multi-domain protein structure prediction; Table S8: The monomer protein dataset from CASP14 used in our benchmark tests. Refs. [172–189] are cited within the Supplementary Materials.

Author Contributions: Q.W.: conceptualization and writing—original draft preparation; Y.C. (Yihan Chen): writing—original draft preparation; Y.S.: data analysis; Y.C. (Yang Cao): writing—review and editing and funding acquisition; G.H.: writing—review and editing and funding acquisition; W.C.: writing—review and editing and funding acquisition; J.G.: writing—review and editing and funding acquisition; W.Z.: writing—review and editing, image design, and data analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Tianjin, grant number 21JCYBJC00340; the National Natural Science Foundation of China, grant number 92370128 and 12326611; and the National Natural Science Foundation of China, grant number 81973243.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation (ACI-1548562). J.G. acknowledges the support of the Chern Institute of Mathematics (CIM) Visiting Scholars Program.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Abbreviations

3D, three-dimensional; AI, artificial intelligence; AlphaFold DB, AlphaFold Protein Structure Database; API, application programming interface; CASP, Critical Assessment of Protein Structure Prediction; CNN, convolutional neural network; DCA, direct coupling analysis; EMBL-EBI, EMBL-European Bioinformatics Institute; Fe-S, iron–sulfur; FM, template-free modeling; GDT, global distance test; HMM, hidden Markov model; MD, molecular dynamics; MSA, multiple sequence alignment; PDB, Protein Data Bank; PLM, protein language model; PSSM, position-specific score matrix; PTM, post-translation modification; REMC, Replica Exchange Monte Carlo; ResNet, deep residual network; TBM, template-based modeling; Zn, zinc.

References

1. Anfinsen, C.B. Principles that Govern the Folding of Protein Chains. *Science* **1973**, *181*, 223–230. [CrossRef]
2. Sanger, F.; Nicklen, S.; Coulson, A.R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463–5467. [CrossRef]
3. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; et al. The Sequence of the Human Genome. *Science* **2001**, *291*, 1304–1351. [CrossRef]
4. Metzker, M.L. Sequencing technologies—The next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46. [CrossRef] [PubMed]
5. Bairoch, A.; Apweiler, R.; Wu, C.H.; Barker, W.C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; et al. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2005**, *33* (Suppl. 1), D154–D159. [CrossRef] [PubMed]
6. Glusker, J.P. X-ray crystallography of proteins. *Methods Biochem. Anal.* **1994**, *37*, 1–72. [CrossRef] [PubMed]
7. Cavanagh, J. *Protein NMR Spectroscopy: Principles and Practice*; Academic Press: Cambridge, MA, USA, 1996.
8. Cheng, Y. Single-Particle Cryo-EM at Crystallographic Resolution. *Cell* **2015**, *161*, 450–457. [CrossRef] [PubMed]
9. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [CrossRef] [PubMed]
10. The UniProt Consortium. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2008**, *36* (Suppl. 1), D190–D195. [CrossRef]
11. Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253*, 694–698. [CrossRef]
12. Lewis, P.N.; Momany, F.A.; Scheraga, H.A. Folding of Polypeptide Chains in Proteins: A Proposed Mechanism for Folding. *Proc. Natl. Acad. Sci. USA* **1971**, *68*, 2293–2297. [CrossRef]
13. McCammon, J.A.; Gelin, B.R.; Karplus, M. Dynamics of folded proteins. *Nature* **1977**, *267*, 585–590. [CrossRef]
14. Bowie, J.U.; Lüthy, R.; Eisenberg, D. A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure. *Science* **1991**, *253*, 164–170. [CrossRef]
15. Skolnick, J.; Kolinski, A. Simulations of the Folding of a Globular Protein. *Science* **1990**, *250*, 1121–1125. [CrossRef]
16. Šali, A.; Blundell, T.L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779–815. [CrossRef]
17. Simons, K.T.; Kooperberg, C.; Huang, E.; Baker, D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *J. Mol. Biol.* **1997**, *268*, 209–225. [CrossRef]
18. Roy, A.; Kucukural, A.; Zhang, Y. I-TASSER: A unified platform for automated protein structure and function prediction. *Nat. Protoc.* **2010**, *5*, 725–738. [CrossRef]
19. Xu, D.; Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins Struct. Funct. Bioinform.* **2012**, *80*, 1715–1735. [CrossRef] [PubMed]
20. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **2015**, *12*, 7–8. [CrossRef] [PubMed]
21. Ovchinnikov, S.; Park, H.; Varghese, N.; Huang, P.-S.; Pavlopoulos, G.A.; Kim, D.E.; Kamisetty, H.; Kyrpides, N.C.; Baker, D. Protein structure determination using metagenome sequence data. *Science* **2017**, *355*, 294–298. [CrossRef] [PubMed]
22. Wang, S.; Sun, S.; Li, Z.; Zhang, R.; Xu, J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **2017**, *13*, e1005324. [CrossRef]
23. Zheng, W.; Li, Y.; Zhang, C.; Pearce, R.; Mortuza, S.M.; Zhang, Y. Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1149–1164. [CrossRef] [PubMed]
24. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710. [CrossRef]
25. Yang, J.; Anishchenko, I.; Park, H.; Peng, Z.; Ovchinnikov, S.; Baker, D. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 1496–1503. [CrossRef]
26. Fischer, D.; Eisenberg, D. Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 11929–11934. [CrossRef] [PubMed]

27. Sánchez, R.; Šali, A. Evaluation of comparative protein structure modeling by MODELLER-3. *Proteins Struct. Funct. Bioinform.* **1997**, *29*, 50–58. [CrossRef]
28. Zhang, Y.; Skolnick, J. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7594–7599. [CrossRef]
29. Malmström, L.; Riffle, M.; Strauss, C.E.M.; Chivian, D.; Davis, T.N.; Bonneau, R.; Baker, D. Superfamily Assignments for the Yeast Proteome through Integration of Structure Prediction with the Gene Ontology. *PLoS Biol.* **2007**, *5*, e76. [CrossRef]
30. Mukherjee, S.; Szilagyi, A.; Roy, A.; Zhang, Y. Genome-Wide Protein Structure Prediction. In *Multiscale Approaches to Protein Modeling: Structure Prediction, Dynamics, Thermodynamics and Macromolecular Assemblies*; Kolinski, A., Ed.; Springer: New York, NY, USA, 2011; pp. 255–279.
31. Xu, D.; Zhang, Y. Ab Initio structure prediction for Escherichia coli: Towards genome-wide protein structure modeling and fold assignment. *Sci. Rep.* **2013**, *3*, 1895. [CrossRef]
32. Zhang, C.; Zheng, W.; Cheng, M.; Omenn, G.S.; Freddolino, P.L.; Zhang, Y. Functions of Essential Genes and a Scale-Free Protein Interaction Network Revealed by Structure-Based Function and Interaction Prediction for a Minimal Genome. *J. Proteome Res.* **2021**, *20*, 1178–1189. [CrossRef]
33. Kim, D.E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **2004**, *32* (Suppl. 2), W526–W531. [CrossRef]
34. Kelley, L.A.; Sternberg, M.J.E. Protein structure prediction on the Web: A case study using the Phyre server. *Nat. Protoc.* **2009**, *4*, 363–371. [CrossRef]
35. Schwede, T.; Kopp, J.R.; Guex, N.; Peitsch, M.C. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **2003**, *31*, 3381–3385. [CrossRef]
36. Söding, J.; Biegert, A.; Lupas, A.N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **2005**, *33* (Suppl. 2), W244–W248. [CrossRef] [PubMed]
37. Wang, Z.; Eickholt, J.; Cheng, J. MULTICOM: A multi-level combination approach to protein structure prediction and its assessments in CASP8. *Bioinformatics* **2010**, *26*, 882–888. [CrossRef] [PubMed]
38. Källberg, M.; Wang, H.; Wang, S.; Peng, J.; Wang, Z.; Lu, H.; Xu, J. Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* **2012**, *7*, 1511–1522. [CrossRef]
39. Xu, J. Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 16856–16865. [CrossRef]
40. Vaidehi, N.; Floriano, W.B.; Trabanino, R.; Hall, S.E.; Freddolino, P.; Choi, E.J.; Zamanakos, G.; Goddard, W.A. Prediction of structure and function of G protein-coupled receptors. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12622–12627. [CrossRef]
41. Zhang, Y.; Thiele, I.; Weekes, D.; Li, Z.; Jaroszewski, L.; Ginalski, K.; Deacon, A.M.; Wooley, J.; Lesley, S.A.; Wilson, I.A.; et al. Three-Dimensional Structural View of the Central Metabolic Network of Thermotoga maritima. *Science* **2009**, *325*, 1544–1549. [CrossRef] [PubMed]
42. Loewenstein, Y.; Raimondo, D.; Redfern, O.C.; Watson, J.; Frishman, D.; Linial, M.; Orengo, C.; Thornton, J.; Tramontano, A. Protein function annotation by homology-based inference. *Genome Biol.* **2009**, *10*, 207. [CrossRef]
43. Radivojac, P.; Clark, W.T.; Oron, T.R.; Schnoes, A.M.; Wittkop, T.; Sokolov, A.; Graim, K.; Funk, C.; Verspoor, K.; Ben-Hur, A.; et al. A large-scale evaluation of computational protein function prediction. *Nat. Methods* **2013**, *10*, 221–227. [CrossRef]
44. Zhang, C.; Zheng, W.; Huang, X.; Bell, E.W.; Zhou, X.; Zhang, Y. Protein Structure and Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as Its Intermediate Host and the Unique Similarity between Its Spike Protein Insertions and HIV-1. *J. Proteome Res.* **2020**, *19*, 1351–1360. [CrossRef]
45. Capriotti, E.; Fariselli, P.; Casadio, R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **2005**, *33* (Suppl. 2), W306–W310. [CrossRef]
46. Tokuriki, N.; Tawfik, D.S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **2009**, *19*, 596–604. [CrossRef]
47. Quan, L.; Lv, Q.; Zhang, Y. STRUM: Structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics* **2016**, *32*, 2936–2946. [CrossRef] [PubMed]
48. Porta-Pardo, E.; Hrabe, T.; Godzik, A. Cancer3D: Understanding cancer mutations through protein structures. *Nucleic Acids Res.* **2015**, *43*, D968–D973. [CrossRef] [PubMed]
49. Pires, D.E.V.; Ascher, D.B.; Blundell, T.L. mCSM: Predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* **2014**, *30*, 335–342. [CrossRef] [PubMed]
50. Porta-Pardo, E.; Godzik, A. Mutation Drivers of Immunological Responses to Cancer. *Cancer Immunol. Res.* **2016**, *4*, 789–798. [CrossRef] [PubMed]
51. Sundaram, L.; Gao, H.; Padigepati, S.R.; McRae, J.F.; Li, Y.; Kosmicki, J.A.; Fritzilas, N.; Hakenberg, J.; Dutta, A.; Shon, J.; et al. Predicting the clinical impact of human mutation with deep neural networks. *Nat. Genet.* **2018**, *50*, 1161–1170. [CrossRef]
52. Woodard, J.; Zhang, C.; Zhang, Y. ADDRESS: A Database of Disease-associated Human Variants Incorporating Protein Structure and Folding Stabilities. *J. Mol. Biol.* **2021**, *433*, 166840. [CrossRef] [PubMed]
53. Evers, A.; Klebe, G. Successful Virtual Screening for a Submicromolar Antagonist of the Neurokinin-1 Receptor Based on a Ligand-Supported Homology Model. *J. Med. Chem.* **2004**, *47*, 5381–5392. [CrossRef]
54. Klebe, G. Virtual ligand screening: Strategies, perspectives and limitations. *Drug Discov. Today* **2006**, *11*, 580–594. [CrossRef]

55. Zhou, H.; Skolnick, J. FINDSITEX: A Structure-Based, Small Molecule Virtual Screening Approach with Application to All Identified Human GPCRs. *Mol. Pharm.* **2012**, *9*, 1775–1784. [CrossRef]
56. Roy, A.; Zhang, Y. Recognizing Protein-Ligand Binding Sites by Global Structural Alignment and Local Geometry Refinement. *Structure* **2012**, *20*, 987–997. [CrossRef] [PubMed]
57. Vajda, S.; Guarnieri, F. Characterization of protein-ligand interaction sites using experimental and computational methods. *Curr. Opin. Drug Discov. Dev.* **2006**, *9*, 354–362.
58. Choudhary, S.; Malik, Y.S.; Tomar, S. Identification of SARS-CoV-2 Cell Entry Inhibitors by Drug Repurposing Using in silico Structure-Based Virtual Screening Approach. *Front. Immunol.* **2020**, *11*, 1664. [CrossRef] [PubMed]
59. Chan, W.K.B.; Zhang, Y. Virtual Screening of Human Class-A GPCRs Using Ligand Profiles Built on Multiple Ligand–Receptor Interactions. *J. Mol. Biol.* **2020**, *432*, 4872–4890. [CrossRef] [PubMed]
60. Kuntz, I.D. Structure-Based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078–1082. [CrossRef] [PubMed]
61. Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960–1964. [CrossRef] [PubMed]
62. Evers, A.; Klabunde, T. Structure-based Drug Discovery Using GPCR Homology Modeling: Successful Virtual Screening for Antagonists of the Alpha1A Adrenergic Receptor. *J. Med. Chem.* **2005**, *48*, 1088–1097. [CrossRef] [PubMed]
63. Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: Applications to targets and beyond. *Br. J. Pharmacol.* **2007**, *152*, 21–37. [CrossRef]
64. Shan, Y.; Kim, E.T.; Eastwood, M.P.; Dror, R.O.; Seeliger, M.A.; Shaw, D.E. How Does a Drug Molecule Find Its Target Binding Site? *J. Am. Chem. Soc.* **2011**, *133*, 9181–9183. [CrossRef] [PubMed]
65. Han, X.; Wang, C.; Qin, C.; Xiang, W.; Fernandez-Salas, E.; Yang, C.-Y.; Wang, M.; Zhao, L.; Xu, T.; Chinnaswamy, K.; et al. Discovery of ARD-69 as a Highly Potent Proteolysis Targeting Chimera (PROTAC) Degrader of Androgen Receptor (AR) for the Treatment of Prostate Cancer. *J. Med. Chem.* **2019**, *62*, 941–964. [CrossRef]
66. Pearce, R.; Zhang, Y. Toward the solution of the protein structure prediction problem. *J. Biol. Chem.* **2021**, *297*, 100870. [CrossRef] [PubMed]
67. Needleman, S.B.; Wunsch, C.D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443–453. [CrossRef]
68. Smith, T.F.; Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147*, 195–197. [CrossRef]
69. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef] [PubMed]
70. Wu, S.; Zhang, Y. MUSTER: Improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins Struct. Funct. Bioinform.* **2008**, *72*, 547–556. [CrossRef]
71. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755–763. [CrossRef]
72. Söding, J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* **2005**, *21*, 951–960. [CrossRef]
73. Buchan, D.W.A.; Jones, D.T. EigenTHREADER: Analogous protein fold recognition by efficient contact map threading. *Bioinformatics* **2017**, *33*, 2684–2690. [CrossRef]
74. Zheng, W.; Wuyun, Q.; Li, Y.; Mortuza, S.M.; Zhang, C.; Pearce, R.; Ruan, J.; Zhang, Y. Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLoS Comput. Biol.* **2019**, *15*, e1007411. [CrossRef]
75. Zhu, J.; Wang, S.; Bu, D.; Xu, J. Protein threading using residue co-variation and deep learning. *Bioinformatics* **2018**, *34*, i263–i273. [CrossRef] [PubMed]
76. Bhattacharya, S.; Roche, R.; Moussad, B.; Bhattacharya, D. DisCovER: Distance- and orientation-based covariational threading for weakly homologous proteins. *Proteins Struct. Funct. Bioinform.* **2022**, *90*, 579–588. [CrossRef] [PubMed]
77. Zheng, W.; Wuyun, Q.; Zhou, X.; Li, Y.; Freddolino, P.L.; Zhang, Y. LOMETS3: Integrating deep learning and profile alignment for advanced protein template recognition and function annotation. *Nucleic Acids Res.* **2022**, *50*, W454–W464. [CrossRef] [PubMed]
78. Zheng, W.; Zhang, C.; Wuyun, Q.; Pearce, R.; Li, Y.; Zhang, Y. LOMETS2: Improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. *Nucleic Acids Res.* **2019**, *47*, W429–W436. [CrossRef] [PubMed]
79. Wu, S.; Zhang, Y. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* **2007**, *35*, 3375–3382. [CrossRef] [PubMed]
80. Zhang, C.; Zheng, W.; Mortuza, S.M.; Li, Y.; Zhang, Y. DeepMSA: Constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. *Bioinformatics* **2020**, *36*, 2105–2112. [CrossRef] [PubMed]
81. Zhang, H.; Shen, Y. Template-based prediction of protein structure with deep learning. *BMC Genom.* **2020**, *21*, 878. [CrossRef]
82. Gao, M.; Skolnick, J. A novel sequence alignment algorithm based on deep learning of the protein folding code. *Bioinformatics* **2021**, *37*, 490–496. [CrossRef]
83. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
84. Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 7112–7127. [CrossRef] [PubMed]
85. Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379*, 1123–1130. [CrossRef] [PubMed]

86. Kaminski, K.; Ludwiczak, J.; Pawlicki, K.; Alva, V.; Dunin-Horkawicz, S. pLM-BLAST: Distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics* **2023**, *39*, btad579. [CrossRef] [PubMed]
87. Pantolini, L.; Studer, G.; Pereira, J.; Durairaj, J.; Tauriello, G.; Schwede, T. Embedding-based alignment: Combining protein language models with dynamic programming alignment to detect structural similarities in the twilight-zone. *Bioinformatics* **2024**, *40*, btad786. [CrossRef] [PubMed]
88. Llinares-López, F.; Berthet, Q.; Blondel, M.; Teboul, O.; Vert, J.-P. Deep embedding and alignment of protein sequences. *Nat. Methods* **2023**, *20*, 104–111. [CrossRef] [PubMed]
89. James, T.M.; Charlie, E.M.S.; Robert, B.; Daniel, B.; Vladimir, G.; Richard, B. Protein Structural Alignments From Sequence. *bioRxiv* **2020**. [CrossRef]
90. Meier, A.; Söding, J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. *PLoS Comput. Biol.* **2015**, *11*, e1004343. [CrossRef]
91. Zheng, W.; Zhang, C.; Bell, E.W.; Zhang, Y. I-TASSER gateway: A protein structure and function prediction server powered by XSEDE. *Future Gener. Comput. Syst.* **2019**, *99*, 73–85. [CrossRef]
92. Yang, J.; Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res.* **2015**, *43*, W174–W181. [CrossRef]
93. Zhang, Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins Struct. Funct. Bioinform.* **2007**, *69*, 108–117. [CrossRef] [PubMed]
94. Song, Y.; DiMaio, F.; Wang, R.Y.-R.; Kim, D.; Miles, C.; Brunette, T.J.; Thompson, J.; Baker, D. High-Resolution Comparative Modeling with RosettaCM. *Structure* **2013**, *21*, 1735–1742. [CrossRef] [PubMed]
95. Piana, S.; Klepeis, J.L.; Shaw, D.E. Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98–105. [CrossRef] [PubMed]
96. Bowie, J.U.; Eisenberg, D. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 4436–4440. [CrossRef] [PubMed]
97. Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins Struct. Funct. Bioinform.* **1994**, *18*, 309–317. [CrossRef] [PubMed]
98. Thomas, D.J.; Casari, G.; Sander, C. The prediction of protein contacts from multiple sequence alignments. *Protein Eng. Des. Sel.* **1996**, *9*, 941–948. [CrossRef] [PubMed]
99. Chiu, D.K.Y.; Kolodziejczak, T. Inferring consensus structure from nucleic acid sequences. *Bioinformatics* **1991**, *7*, 347–352. [CrossRef] [PubMed]
100. Li, Y.; Zhang, C.; Bell, E.W.; Yu, D.-J.; Zhang, Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1082–1091. [CrossRef]
101. Li, Y.; Zhang, C.; Bell, E.W.; Zheng, W.; Zhou, X.; Yu, D.-J.; Zhang, Y. Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. *PLoS Comput. Biol.* **2021**, *17*, e1008865. [CrossRef]
102. Adhikari, B.; Cheng, J. CONFOLD2: Improved contact-driven ab initio protein structure modeling. *BMC Bioinform.* **2018**, *19*, 22. [CrossRef]
103. Li, Y.; Hu, J.; Zhang, C.; Yu, D.-J.; Zhang, Y. ResPRE: High-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. *Bioinformatics* **2019**, *35*, 4647–4655. [CrossRef] [PubMed]
104. Ding, W.; Gong, H. Predicting the Real-Valued Inter-Residue Distances for Proteins. *Adv. Sci.* **2020**, *7*, 2001314. [CrossRef] [PubMed]
105. Greener, J.G.; Kandathil, S.M.; Jones, D.T. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nat. Commun.* **2019**, *10*, 3977. [CrossRef] [PubMed]
106. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.R.; Bridgland, A.; et al. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1141–1148. [CrossRef] [PubMed]
107. Du, Z.; Su, H.; Wang, W.; Ye, L.; Wei, H.; Peng, Z.; Anishchenko, I.; Baker, D.; Yang, J. The trRosetta server for fast and accurate protein structure prediction. *Nat. Protoc.* **2021**, *16*, 5634–5651. [CrossRef]
108. Zheng, W.; Li, Y.; Zhang, C.; Zhou, X.; Pearce, R.; Bell, E.W.; Huang, X.; Zhang, Y. Protein structure prediction using deep learning distance and hydrogen-bonding restraints in CASP14. *Proteins Struct. Funct. Bioinform.* **2021**, *89*, 1734–1751. [CrossRef] [PubMed]
109. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **2020**, *588*, 203–204. [CrossRef]
110. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
111. Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making protein folding accessible to all. *Nat. Methods* **2022**, *19*, 679–682. [CrossRef]
112. Gustaf, A.; Nazim, B.; Christina, F.; Sachin, K.; Qinghui, X.; William, G.; Timothy, J.O.D.; Daniel, B.; Ian, F.; Niccolò, Z.; et al. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *bioRxiv* **2023**. [CrossRef]
113. Ziyao, L.; Xuyang, L.; Weijie, C.; Fan, S.; Hangrui, B.; Guolin, K.; Linfeng, Z. Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. *bioRxiv* **2022**. [CrossRef]

114. Steinegger, M.; Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [CrossRef]
115. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
116. Pearce, R.; Zhang, Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr. Opin. Struct. Biol.* **2021**, *68*, 194–207. [CrossRef] [PubMed]
117. Xu, J.; Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **2010**, *26*, 889–895. [CrossRef] [PubMed]
118. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Wang, J.; Cong, Q.; Kinch, L.N.; Schaeffer, R.D.; et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. [CrossRef]
119. Fang, X.; Wang, F.; Liu, L.; He, J.; Lin, D.; Xiang, Y.; Zhu, K.; Zhang, X.; Wu, H.; Li, H.; et al. A method for multiple-sequence-alignment-free protein structure prediction using a protein language model. *Nat. Mach. Intell.* **2023**, *5*, 1087–1096. [CrossRef]
120. Jin, S.; Chenchen, H.; Yuyang, Z.; Junjie, S.; Xibin, Z.; Fajie, Y. SaProt: Protein Language Modeling with Structure-aware Vocabulary. *bioRxiv* **2023**. [CrossRef]
121. Ruidong, W.; Fan, D.; Rui, W.; Rui, S.; Xiwen, Z.; Shitong, L.; Chenpeng, S.; Zuofan, W.; Qi, X.; Bonnie, B.; et al. High-resolution *de novo* structure prediction from primary sequence. *bioRxiv* **2022**. [CrossRef]
122. Konstantin, W.; Michael, H.; Martin, S.; Burkhard, R. Ultra-fast protein structure prediction to capture effects of sequence variation in mutation movies. *bioRxiv* **2022**. [CrossRef]
123. Schauperl, M.; Denny, R.A. AI-Based Protein Structure Prediction in Drug Discovery: Impacts and Challenges. *J. Chem. Inf. Model.* **2022**, *62*, 3142–3156. [CrossRef]
124. Chothia, C.; Gough, J.; Vogel, C.; Teichmann, S.A. Evolution of the Protein Repertoire. *Science* **2003**, *300*, 1701–1703. [CrossRef]
125. Wollacott, A.M.; Zanghellini, A.; Murphy, P.; Baker, D. Prediction of structures of multidomain proteins from structures of the individual domains. *Protein Sci.* **2007**, *16*, 165–175. [CrossRef]
126. Xu, D.; Jaroszewski, L.; Li, Z.; Godzik, A. AIDA: Ab initio domain assembly for automated multi-domain protein structure prediction and domain–domain interaction prediction. *Bioinformatics* **2015**, *31*, 2098–2105. [CrossRef]
127. Zhou, X.; Peng, C.; Zheng, W.; Li, Y.; Zhang, G.; Zhang, Y. DEMO2: Assemble multi-domain protein structures by coupling analogous template alignments with deep-learning inter-domain restraint prediction. *Nucleic Acids Res.* **2022**, *50*, W235–W245. [CrossRef] [PubMed]
128. Zhou, X.; Hu, J.; Zhang, C.; Zhang, G.; Zhang, Y. Assembling multidomain protein structures through analogous global structural alignments. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 15930–15938. [CrossRef]
129. Peng, C.-X.; Zhou, X.-G.; Xia, Y.-H.; Liu, J.; Hou, M.-H.; Zhang, G.-J. Structural analogue-based protein structure domain assembly assisted by deep learning. *Bioinformatics* **2022**, *38*, 4513–4521. [CrossRef] [PubMed]
130. Zhou, X.; Zheng, W.; Li, Y.; Pearce, R.; Zhang, C.; Bell, E.W.; Zhang, G.; Zhang, Y. I-TASSER-MTD: A deep-learning-based platform for multi-domain protein structure and function prediction. *Nat. Protoc.* **2022**, *17*, 2326–2353. [CrossRef] [PubMed]
131. Zheng, W.; Zhou, X.; Wuyun, Q.; Pearce, R.; Li, Y.; Zhang, Y. FUpred: Detecting protein domains through deep-learning-based contact map prediction. *Bioinformatics* **2020**, *36*, 3749–3757. [CrossRef] [PubMed]
132. Xue, Z.; Xu, D.; Wang, Y.; Zhang, Y. ThreaDom: Extracting protein domain boundary information from multiple threading alignments. *Bioinformatics* **2013**, *29*, i247–i256. [CrossRef] [PubMed]
133. Zheng, W.; Wuyun, Q.; Freddolino, P.L.; Zhang, Y. Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1684–1703. [CrossRef]
134. Xia, Y.; Zhao, K.; Liu, D.; Zhou, X.; Zhang, G. Multi-domain and complex protein structure prediction using inter-domain interactions from deep learning. *Commun. Biol.* **2023**, *6*, 1221. [CrossRef]
135. Zhu, H.-T.; Xia, Y.-H.; Zhang, G.-J. E2EDA: Protein Domain Assembly Based on End-to-End Deep Learning. *J. Chem. Inf. Model.* **2023**, *63*, 6451–6461. [CrossRef]
136. Li, Y.; Zhang, C.; Yu, D.-J.; Zhang, Y. Deep learning geometrical potential for high-accuracy *ab initio* protein structure prediction. *iScience* **2022**, *25*, 104425. [CrossRef] [PubMed]
137. Moult, J.; Pedersen, J.T.; Judson, R.; Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins Struct. Funct. Bioinform.* **1995**, *23*, ii-iv. [CrossRef]
138. Zhang, Y.; Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins Struct. Funct. Bioinform.* **2004**, *57*, 702–710. [CrossRef] [PubMed]
139. Simpkin, A.J.; Mesdaghi, S.; Sánchez Rodríguez, F.; Elliott, L.; Murphy, D.L.; Kryshtafovych, A.; Keegan, R.M.; Rigden, D.J. Tertiary structure assessment at CASP15. *Proteins* **2023**, *91*, 1616–1635. [CrossRef]
140. Robin, X.; Studer, G.; Durairaj, J.; Eberhardt, J.; Schwede, T.; Walters, W.P. Assessment of protein–ligand complexes in CASP15. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1811–1821. [CrossRef]
141. Pang, M.; He, W.; Lu, X.; She, Y.; Xie, L.; Kong, R.; Chang, S. CoDock-Ligand: Combined template-based docking and CNN-based scoring in ligand binding prediction. *BMC Bioinform.* **2023**, *24*, 444. [CrossRef]

142. Xu, X.; Duan, R.; Zou, X. Template-guided method for protein–ligand complex structure prediction: Application to CASP15 protein–ligand studies. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1829–1836. [CrossRef]
143. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [CrossRef]
144. Liu, X.; Jiang, H.; Li, H. SHAFTS: A Hybrid Approach for 3D Molecular Similarity Calculation. 1. Method and Assessment of Virtual Screening. *J. Chem. Inf. Model.* **2011**, *51*, 2372–2385. [CrossRef]
145. Shen, T.; Liu, F.; Wang, Z.; Sun, J.; Bu, Y.; Meng, J.; Chen, W.; Yao, K.; Mu, Y.; Li, W.; et al. zPoseScore model for accurate and robust protein–ligand docking pose scoring in CASP15. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1837–1849. [CrossRef]
146. Kotelnikov, S.; Ashizawa, R.; Popov, K.I.; Khan, O.; Ignatov, M.; Li, S.X.; Hassan, M.; Coutsias, E.A.; Poda, G.; Padhorny, D.; et al. Accurate ligand–protein docking in CASP15 using the ClusPro LigTBM server. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1822–1828. [CrossRef]
147. Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749. [CrossRef]
148. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439–D444. [CrossRef] [PubMed]
149. The UniProt, C. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489. [CrossRef] [PubMed]
150. Hekkelman, M.L.; de Vries, I.; Joosten, R.P.; Perrakis, A. AlphaFill: Enriching AlphaFold models with ligands and cofactors. *Nat. Methods* **2023**, *20*, 205–213. [CrossRef]
151. Wehrspan, Z.J.; McDonnell, R.T.; Elcock, A.H. Identification of Iron-Sulfur (Fe-S) Cluster and Zinc (Zn) Binding Sites Within Proteomes Predicted by DeepMind's AlphaFold2 Program Dramatically Expands the Metalloproteome. *J. Mol. Biol.* **2022**, *434*, 167377. [CrossRef]
152. Jakubec, D.; Skoda, P.; Krivak, R.; Novotny, M.; Hoksza, D. PrankWeb 3: Accelerated ligand-binding site predictions for experimental and modelled protein structures. *Nucleic Acids Res.* **2022**, *50*, W593–W597. [CrossRef]
153. Bludau, I.; Willems, S.; Zeng, W.-F.; Strauss, M.T.; Hansen, F.M.; Tanzer, M.C.; Karayel, O.; Schulman, B.A.; Mann, M. The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol.* **2022**, *20*, e3001636. [CrossRef] [PubMed]
154. van Kempen, M.; Kim, S.S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C.L.M.; Söding, J.; Steinegger, M. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **2023**. [CrossRef] [PubMed]
155. Aderinwale, T.; Bharadwaj, V.; Christoffer, C.; Terashi, G.; Zhang, Z.; Jahandideh, R.; Kagaya, Y.; Kihara, D. Real-time structure search and structure classification for AlphaFold protein models. *Commun. Biol.* **2022**, *5*, 316. [CrossRef] [PubMed]
156. Bordin, N.; Sillitoe, I.; Nallapareddy, V.; Rauer, C.; Lam, S.D.; Waman, V.P.; Sen, N.; Heinzinger, M.; Littmann, M.; Kim, S.; et al. AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun. Biol.* **2023**, *6*, 160. [CrossRef] [PubMed]
157. David, A.; Islam, S.; Tankhilevich, E.; Sternberg, M.J.E. The AlphaFold Database of Protein Structures: A Biologist's Guide. *J. Mol. Biol.* **2022**, *434*, 167336. [CrossRef] [PubMed]
158. Hou, M.; Jin, S.; Cui, X.; Peng, C.; Zhao, K.; Song, L.; Zhang, G. Protein Multiple Conformation Prediction Using Multi-Objective Evolution Algorithm. *Interdiscip. Sci. Comput. Life Sci.* **2024**. [CrossRef]
159. Wayment-Steele, H.K.; Ojoawo, A.; Otten, R.; Apitz, J.M.; Pitsawong, W.; Hömberger, M.; Ovchinnikov, S.; Colwell, L.; Kern, D. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **2023**. [CrossRef] [PubMed]
160. del Alamo, D.; Sala, D.; McHaourab, H.S.; Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **2022**, *11*, e75751. [CrossRef] [PubMed]
161. Park, S.H.; Ayoub, A.; Lee, Y.-T.; Xu, J.; Kim, H.; Zheng, W.; Zhang, B.; Sha, L.; An, S.; Zhang, Y.; et al. Cryo-EM structure of the human MLL1 core complex bound to the nucleosome. *Nat. Commun.* **2019**, *10*, 5540. [CrossRef]
162. Lee, Y.-T.; Ayoub, A.; Park, S.-H.; Sha, L.; Xu, J.; Mao, F.; Zheng, W.; Zhang, Y.; Cho, U.-S.; Dou, Y. Mechanism for DPY30 and ASH2L intrinsically disordered regions to modulate the MLL/SET1 activity on chromatin. *Nat. Commun.* **2021**, *12*, 2953. [CrossRef]
163. Zhang, H.; Shang, R.; Kim, K.; Zheng, W.; Johnson, C.J.; Sun, L.; Niu, X.; Liu, L.; Zhou, J.; Liu, L.; et al. Evolution of a chordate-specific mechanism for myoblast fusion. *Sci. Adv.* **2022**, *8*, eadd2696. [CrossRef]
164. Wu, S.; Tian, C.; Liu, P.; Guo, D.; Zheng, W.; Huang, X.; Zhang, Y.; Liu, L. Effects of SARS-CoV-2 mutations on protein structures and intraviral protein–protein interactions. *J. Med. Virol.* **2021**, *93*, 2132–2140. [CrossRef]
165. Zheng, W.; Zhang, C.; Li, Y.; Pearce, R.; Bell, E.W.; Zhang, Y. Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. *Cell Rep Methods* **2021**, *1*, 100014. [CrossRef] [PubMed]
166. Richard, E.; Michael, O.N.; Alexander, P.; Natasha, A.; Andrew, S.; Tim, G.; Augustin, Ž.; Russ, B.; Sam, B.; Jason, Y.; et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* **2022**. [CrossRef]
167. Chen, K.; Zhou, Y.; Wang, S.; Xiong, P. RNA tertiary structure modeling with BRiQ potential in CASP15. *Proteins Struct. Funct. Bioinform.* **2023**, *91*, 1771–1778. [CrossRef] [PubMed]

168. Li, Y.; Zhang, C.; Feng, C.; Pearce, R.; Lydia Freddolino, P.; Zhang, Y. Integrating end-to-end learning with deep geometrical potentials for ab initio RNA structure prediction. *Nat. Commun.* **2023**, *14*, 5745. [CrossRef] [PubMed]
169. Wang, W.; Feng, C.; Han, R.; Wang, Z.; Ye, L.; Du, Z.; Wei, H.; Zhang, F.; Peng, Z.; Yang, J. trRosettaRNA: Automated prediction of RNA 3D structure with transformer network. *Nat. Commun.* **2023**, *14*, 7266. [CrossRef]
170. Baek, M.; McHugh, R.; Anishchenko, I.; Jiang, H.; Baker, D.; DiMaio, F. Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **2024**, *21*, 117–121. [CrossRef]
171. Terwilliger, T.C.; Liebschner, D.; Croll, T.I.; Williams, C.J.; McCoy, A.J.; Poon, B.K.; Afonine, P.V.; Oeffner, R.D.; Richardson, J.S.; Read, R.J.; et al. AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat. Methods* **2024**, *21*, 110–116. [CrossRef]
172. Xu, D.; Jaroszewski, L.; Li, Z.; Godzik, A. FFAS-3D: Improving fold recognition by including optimized structural features and template re-ranking. *Bioinformatics* **2014**, *30*, 660–667. [CrossRef] [PubMed]
173. Ma, J.; Wang, S.; Wang, Z.; Xu, J. MRFalign: Protein Homology Detection through Alignment of Markov Random Fields. *PLoS Comput. Biol.* **2014**, *10*, e1003500. [CrossRef]
174. Cheng, J.; Li, J.; Wang, Z.; Eickholt, J.; Deng, X. The MULTICOM toolbox for protein structure prediction. *BMC Bioinform.* **2012**, *13*, 65. [CrossRef]
175. Cheng, J. A multi-template combination algorithm for protein comparative modeling. *BMC Struct. Biol.* **2008**, *8*, 18. [CrossRef]
176. Kelley, L.A.; Mezulis, S.; Yates, C.M.; Wass, M.N.; Sternberg, M.J.E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **2015**, *10*, 845–858. [CrossRef] [PubMed]
177. Peng, J.; Xu, J. Raptorx: Exploiting structure information for protein alignment by statistical inference. *Proteins Struct. Funct. Bioinform.* **2011**, *79*, 161–171. [CrossRef]
178. Yang, Y.; Faraggi, E.; Zhao, H.; Zhou, Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* **2011**, *27*, 2076–2082. [CrossRef] [PubMed]
179. Jones, D.T. Predicting novel protein folds by using FRAGFOLD. *Proteins Struct. Funct. Bioinform.* **2001**, *45*, 127–132. [CrossRef] [PubMed]
180. Rohl, C.A.; Strauss, C.E.M.; Misura, K.M.S.; Baker, D. Protein Structure Prediction Using Rosetta. In *Methods in Enzymology*; Academic Press: Cambridge, MA, USA, 2004; Volume 383, pp. 66–93.
181. Mortuza, S.M.; Zheng, W.; Zhang, C.; Li, Y.; Pearce, R.; Zhang, Y. Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. *Nat. Commun.* **2021**, *12*, 5011. [CrossRef] [PubMed]
182. Pearce, R.; Li, Y.; Omenn, G.S.; Zhang, Y. Fast and accurate Ab Initio Protein structure prediction using deep learning potentials. *PLoS Comput. Biol.* **2022**, *18*, e1010539. [CrossRef] [PubMed]
183. Shen, T.; Wu, J.; Lan, H.; Zheng, L.; Pei, J.; Wang, S.; Liu, W.; Huang, J. When homologous sequences meet structural decoys: Accurate contact prediction by tFold in CASP14—(tFold for CASP14 contact prediction). *Proteins Struct. Funct. Bioinform.* **2021**, *89*, 1901–1910. [CrossRef]
184. Cheng, S.; Zhao, X.; Lu, G.; Fang, J.; Yu, Z.; Zheng, T.; Wu, R.; Zhang, X.; Peng, J.; You, Y. FastFold: Reducing AlphaFold Training Time from 11 Days to 67 Hours. *arXiv* **2022**, arXiv:2203.00854.
185. Wang, G.; Fang, X.; Wu, Z.; Liu, Y.; Xue, Y.; Xiang, Y.; Yu, D.; Wang, F.; Ma, Y. HelixFold: An Efficient Implementation of AlphaFold2 using PaddlePaddle. *arXiv* **2022**, arXiv:2207.05477.
186. Liu, S.; Zhang, J.; Chu, H.; Wang, M.; Xue, B.; Ni, N.; Yu, J.; Xie, Y.; Chen, Z.; Chen, M.; et al. PSP: Million-level Protein Sequence Dataset for Protein Structure Prediction. *arXiv* **2022**, arXiv:2206.12240.
187. Ruffolo, J.A.; Chu, L.-S.; Mahajan, S.P.; Gray, J.J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat. Commun.* **2023**, *14*, 2389. [CrossRef] [PubMed]
188. Jing, X.; Wu, F.; Luo, X.; Xu, J. RaptorX-Single: Single-sequence protein structure prediction by integrating protein language models. *bioRxiv* **2023**. bioRxiv:2023.04.24.538081.
189. Wang, W.; Peng, Z.; Yang, J. Single-sequence protein structure prediction using supervised transformer protein language models. *Nat. Comput. Sci.* **2022**, *2*, 804–814. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.