Beyond AlphaFold2: the impact of AI for the further improvement of protein structure prediction

Beyond AlphaFold2: the Impact of AI for the Further Improvement of Protein Structure Prediction

Ahmet Gurkan Genc, Liam J. McGuffin

School of Biological Sciences, University of Reading, Reading RG6 6EX, UK

Abstract

Protein structure prediction is fundamental to molecular biology and has numerous applications in areas such as drug discovery and protein engineering. Machine learning techniques have greatly advanced protein 3D modelling in recent years, particularly with the development of AlphaFold2 (AF2), which can analyze sequences of amino acids and predict 3D structures with near experimental accuracy. Since the release of AF2, numerous studies have been conducted, either using AF2 directly for large scale modelling, or building upon the software for other use cases. Many reviews have been published discussing the impact of AF2 in the field of protein bioinformatics, particularly in relation to neural networks, which have highlighted what AF2 can and cannot do. It is evident that AF2 and similar approaches are open to further development and several new approaches have emerged, in addition to older refinement approaches, for improving the quality of predictions. Here we provide a brief overview, aimed at the general biologist, of how machine learning techniques have been used for improvement of 3D models of proteins following AF2, and we highlight the impacts of these approaches. In the most recent experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP15), the most successful groups all developed their own tools for protein structure modelling that were based at least in some part on AF2. This improvement involved employing techniques such as, generative modelling, changing parameters such as dropout to generate more AF2 structures, and data-driven approaches including using alternative templates and MSAs.

Key words Protein refinement, machine learning, data-driven approaches, AlphaFold, CASP

1 Introduction

Proteins are complex molecules that play a vital role in the functioning of living organisms. They are composed of long chains of amino acids, which fold into a specific 3D structure that is determined by the sequence of these amino acids. Determining the 3D structure of a protein is crucial for understanding its function and how it interacts with other molecules [1]. In addition to the ordered structures in protein structure, such as alpha helices, beta sheets and coils [2], there are also intrinsically disordered regions within a protein structure that play a functional role by inducing conformational changes in the 3D structure of a protein. These regions lack a well-defined structure and exhibit flexibility, allowing proteins to adopt different forms to perform alternative functions not based on sequence-structure-function paradigm [3,4]. Furthermore, the presence of intrinsically unstructured or disordered proteins [5], which are associated with diseases like Alzheimer and Parkinson [6], along with the existence of approximately up to 30% disorder-promoting regions in eukaryotic proteomes [7], is widely acknowledged by bioinformaticians in protein modeling. Protein modelling is the process of predicting the 3D structure of a protein based on its amino acid sequence. This can be done using various computational methods, such as homology modelling, which involves using the 3D structure of a similar protein as a template, or de novo modelling, which involves predicting the structure without the use of a template. Once a protein model has been generated, it is often necessary to refine it to correct any errors and improve its accuracy. This process, known as refinement, can involve using experimental data or other predicted information to optimize the model and make it more consistent with known biophysical constraints. Refinement is an important step in protein modelling as it can improve the accuracy and reliability of the model, which is essential for its use in downstream applications such as drug design or protein engineering [1].

2 The History of Protein Modelling

With the advent of modern biochemistry, a significant problem for biochemists in the 1950s was

how to deduce the structural forms and functions of protein molecules. It was known that the protein amino acid sequences, also known as primary structures, govern the unique chemical characteristics of distinct proteins, and most biochemists have guided their work based on this understanding [8]. In light of this, Frederick Sanger was the first person to design an experimental method in 1955 to obtain the protein sequence of insulin [9]. Christian Anfinsen also made significant contributions to our understanding of proteins, developing his theory of protein folding and summarizing in 1972, that the native conformation, or the unique 3D structure of a given protein [10], is determined by the sum of interatomic interactions generated from the information stored in the primary structure of the protein [8].

The goal of protein structure prediction is to determine the Euclidean coordinates of every atom in a protein using computational methods, starting from the protein's amino acid sequence. Traditionally, there have been two main approaches to protein structure prediction: template-based modelling (TBM) and template- free modelling (FM). In TBM, models are constructed using the structures of similar proteins as templates, which are stored in the Protein Data Bank (PDB). FM, on the other hand, seeks to predict protein structures without using templates of known structures. However, recent advances in deep learning have significantly affected the field of protein structure prediction, leading to the development of tools that can accurately predict high-quality geometric properties of proteins, such as inter-residue contacts, distances, and torsion angles. These tools, which combine structural features generated by deep learning with classical folding methods, have greatly improved the quality of FM protein structure prediction, where there are no homologous templates available [11].

Modern prediction systems of protein structure typically consist of four main modules: input, a computational “trunk”, output, and refinement. The input module converts the protein sequence into various forms of information, such as a protein-specific scoring matrix, pairwise interactions (co-evolutionary information), Markov random fields, and a raw multiple sequence alignment. The trunk module, usually a neural network, converts this information into spatial

information about the protein's structure. The output module produces representations of the protein's structure, such as a contact map, distogram, or orientogram. Finally, the refinement module improves the quality of the 3D structure and generates atomic coordinates if needed. Traditionally, these operations have relied on a combination of physics-based energy functions, knowledge-based statistical reasoning, and heuristic algorithms. However, more recently, protein structure prediction tools have started using more sophisticated neural networks and machine learning algorithms in an end-to-end approach, rather than individually modelling each component [12].

The extension of a residual neural network to predict distance information between two residues within a certain range greatly improved residue-residue contact prediction [13]. Using this distance map, generated by 220 residual blocks from a protein sequence, AlphaFold (AF) demonstrated its capabilities in the CASP13 experiment [14]. These predicted distances were then used in structural fragment assembly and folding simulation through gradient descent. AF2, the second version of AF developed by the DeepMind team, made an exciting breakthrough in the field of protein structure prediction in the CASP14 [11]. Its modelled structures achieved unprecedented accuracy, making the longstanding problem of protein folding seem trivial [15].

One of the main differences between the first and second version of AF was in their neural network architectures. While AF used a convolutional neural network and employed distance map prediction, AF2 used an attention-based neural network and took richer input information from the full multiple sequence alignment (MSA). Instead of using gradient descent optimization, as in AF, AF2 evolved a fully end-to-end deep learning algorithm that predicts the 3D structure solely from the sequence, with an iterative refinement- like process, termed recycling, based on its local error estimation score [11]. While there were many AF2 models that reached near experimental accuracy in CASP14, for many targets the AF2 models still include significant errors, and so there is still room for further improvement. To ensure the more active utilization of structure prediction beyond AF2, the importance of refinement will be further emphasized.

3 Methods for the Refinement of 3D Models of Protein Structures before AF2

The objective of computational protein structure refinement is to optimize a starting structure to its native state, such that it approaches the accuracy of experimental measurements. To achieve this, refinement methods of protein structure utilize conformational sampling to optimize an atomic force field to bring the starting structure closer to its native conformation [16]. Two basic concepts have been so far used to obtain and improve the three-dimensional structures of proteins from protein sequences: 1) physical interactions and 2) evolutionary information from bioinformatics analysis. Evolutionary approaches had become more useful because physical interactions, such as thermodynamic or kinetic information from protein physics or statistical approaches, can create a computational burden for molecular simulations [17]. Therefore, tools based on physical laws could only be used for refinement of predicted structures following initial modelling using evolutionary approaches.

Before the development of AF2, refinement methods were generally divided into three categories: molecular dynamics (MD) as shown in Figure 1 and Monte Carlo simulation, energy minimization, and fragment assembly. Molecular simulations and physics-based approaches sample multiple MD trajectories according to the physical principles governing atom-atom interactions. Energy minimization methods aim to find the lowest energy structure by repackaging both backbone and side-chain atoms based on physical and knowledge- based force fields, while fragment-based methods use template fragment information from the PDB in conjunction with statistical potentials. However, these refinement methods require extensive conformational sampling, which can be time-consuming and computationally demanding [18]. A common approach in refinement protocols without the use of machine learning is the application of restraints. Many protocols have adopted this method in different ways. For example, the latest version of the ReFOLD tool- ReFOLD3 [19], uses restraints based on quality estimation scores, while approaches developed by Feig use restraints on backbone atoms [20].

Even prior to the release of AF2, some machine learning approaches had been used for

refinement. RefineD [16] applied a machine learning-based approach for generating restraints to effectively guide conformational sampling. In machine learning, feature extraction is a useful technique for preparing training datasets, which is beneficial in terms of constraining sampling space. However, extracting features from raw data requires manual effort. End-to-end approaches, which do not require feature extraction, outperformed traditional methods such as principal component analysis (PCA). One of the first end-to-end deep learning methods, with feature neutralization, for protein structure prediction was SASNet [21], which developed for protein interface prediction. rawMSA [22] also used an end-to-end method based on MSA embedding as input to a convolutional neural network (CNN). However, the protein refinement area had ignored the concept of feature neuralization until recently. NEMO [23] was the first tool to use neuralization employing an explicit 3D simulator to fold proteins. NEMO`s principle is to build an end-to-end model from sequence by updating to protein position based on Langevin dynamics, which proves to learn itself without co-evolutionary information, leading to competitive results [12]. It is clear that using end-to-end methods like AF2, it is possible to capture more distant evolutionary relationships from MSAs and to explore the protein space more effectively on both a sequence and structural basis, than slower MD-based methods, based on physical laws.

4 Machine Learning Approaches for the Refinement of 3D Models of Protein Structure via AF2

Machine learning-based approaches have had a more significant impact on various aspects of protein research [12] while the deep learning approaches, using different kinds of neural networks have been focused on 3D protein structure refinement in recent years. In the first machine learning-based refinement methods after AF2, such as DeepRefiner [24], models were generated by using deep neural networks to calculate the error rate at the residue level from the initial structures, followed by energy minimization [24]. However, these methods do not always predict with the same accuracy for all residues in the protein because of existing disorder regions in protein structures and the problem of using of limited-size MSA [25]. With the development of deep learning methods, different data other than the sequence or the 3D structure of the protein have been used in training

to refine protein models. For example, cryo-EM density data has been used to ensure that high-quality atom-level protein structures are further refined. In this regard, AlphaFold2-Phenix made use of both Cryo-EM data and deep learning via AF2 [26], which is one of the basic methods for refining AF2 structures.

Thanks to AF2, the significance of backbone refinement has decreased as the backbone accuracies have increased up to 0.96 Å [27]. Rather than backbone refinement, the improvement of protein structure predictions based on the side chains has been a long- standing goal, and the methods commonly used for this purpose are based on energy minimization techniques, rotamer libraries, or energy functions. However, these approaches tend to be weak in detecting long-range interactions. One of the successes of AF2 is that it correctly predicts such interactions by combining residue covariations with deep learning. OPUS-rota4 [28] on the other hand, has been developed to better predict the FM type models predicted by AF2 which is the first side-chain repacking method using machine learning, by taking MSA as an input. This approach utilizes a sequential deep network architecture to predict side-chain coordinates. The method first derives an initial model using side-chain orientation information, and subsequently refines it through a submodule, with gradient descent optimization based on predicted distance constraints, resulting in a final protein structure prediction [29].

Now that tertiary protein structures can be predicted at the near experimental level, the focus has shifted more towards the prediction of protein complexes. Traditional machine learning and deep learning approaches, particularly after the release of AF2, have gained attention for use in modelling protein complexes. Deep learning models have been utilized to aid in the refinement of residue positions within tertiary protein structures or predict refined residue positions using indirect target values, such as inter-residue distances. However, these methods, after refinement, necessitate all-atom restoration procedures as a subsequent step, in order to recover the positions of backbone and side-chain atoms. Deep learning methods were initially demonstrated to be successful for tertiary structures rather than complex structures due to their high computational

memory complexity, which require large datasets for training [30]. Currently, there is no standalone machine learning based tool for refining protein-protein interaction (PPI) models after AF2, and existing standalone PPI refinement methods use a combination of different protocols. While the AF2-Multimer version has been developed to successfully modelled complexes, there is still a gap between the predicted structures and experimental structures. Therefore, refinement approaches are still required for both AF2- Multimer and non-AF2 based models of complexes. One limitation of AF2-Multimer is that it did not predict structures with more than two chains in a protein complex very well. This interaction gap was addressed using the Monte Carlo Tree algorithm, but it was noted that the MolPC [31] method still had a problem with estimating stoichiometry [31].

5 Data-Driven Approaches for the Refinement of 3D Models of Protein Structure

Knowledge-based methods are widely used for the improvement methods for protein modelling as a guide to sampling based on existing data [32]. This information can be derived from experimental methods such as X-ray crystallography and nuclear magnetic resonance [33] or the use of other structural information, e.g. fragments [34]. In recent times, the most widely used experimental data in knowledge-based approaches is cryo-EM data (as mentioned previously), which is integrated with AF2 to improve predictions made by deep learning methods [25]. Luckily, with the increase in data availability, there has been a rapid increase in the use of data- driven methods, and deep generative modelling have gained the great interest in providing new sequence production. One advantage of these methods is that they overcome the limitations of traditional statistical sequence methods used in protein analysis, which are not effective in capturing the relationships between different protein families. This enables the extraction of better evolutionary information from sequences and better definition of intrinsic restraints [35]. Figure 2 demonstrates the workflow of general refinement based on data and the different approaches are detailed in the following subsections.

5.1 Co-Evolution Data-Based Approaches

Multiple Sequence Alignments (MSAs) provide the inputs to extract coevolution information to

machine learning approaches, based on the principle that a residue that is in close contact with another residue may co-mutate with that residue in order to preserve protein structure and function. The more homologous the sequences that there are in an MSA, the more accurately the protein structure can be predicted [36]. A general approach for enhancing the quality of predicted structures from methods such as AF2 can be based on improving the input MSAs. One of the key features that sets AF2 apart from other state-of-the-art methods is the use and inference of MSAs. The success achieved here has inspired the community by creating higher quality (or deeper) MSAs using various deep alignment methods. One of the tools that has gained popularity is MMseq2 [37], which is used by ColabFold [38] and is computationally efficient. Hence, ColabFold often produces deeper MSAs using MMseq2 and, as a result, can be used to generate better structures than DeepMind's AF2 [38].

One alternative method for directly extracting evolutionary information is to use deep learning methods thanks to deep latent representation, which encode information from billions of known protein sequences by adapting language models from natural language processing (NLP). Using generative models of protein sequences has been a successful approach for obtaining deeper evolutionary information and functional protein sequences [39]. In addition to protein sequences, metagenomic data contains a vast amount of information for discovering new proteins with functional structures. As a result, metagenomic data has received particular attention for uncovering evolutionary information on proteins of interest by incorporating it into MSAs, which can be used as input to deep learning tools to improve protein structure quality. While there are many types of metagenomic databases, the combination of genomic and metagenomic databases is becoming increasingly important, as there is limited information on them in UniProtKB [40]. What is significant here is the increase in the Neff (number of effective sequences) value with metagenomic data. However, using more metagenomic data may not always result in a more accurate MSA [41]. Yang et al [42] discovered that using one or a few specific microbes linked to the target protein family can be more beneficial in constructing MSAs. The other significant issue is the technique that is used to

search the metagenomic databases, particularly when used for modelling protein interactions.

Co-evolutionary approaches, such as direct coupling analysis (DCA), can be utilized to predict the interactions between protein chains. Originally, these methods were applied to predict the interactions of bacterial two-component signalling proteins. However, the accuracy of these predictions has since been improved through the incorporation of machine learning techniques. Bryant's et al. [43] emphasized the importance of producing optimal MSAs for generating complex proteins using the Fold and Dock approach, which utilizes distance and angle information as constraints. In this approach, the sequence for each chain is first classified according to their taxonomy and matching was achieved between the sequences with the highest number of hits for same organism in each chain. This customised MSA approach led to improvements in the Dock-Q score [44] indicating improved model quality [43].

Due to its reliance on co-evolutionary information derived from multiple sequence alignments, AF2 demonstrates suboptimal performance when applied to proteins that lack homologous sequences. This is projected to include 20% of all metagenomic protein sequences and 11% of eukaryotic and viral proteins. In addition, the integration of MSAs with deep learning has been shown to be weak in terms of the antibody-antigen model of the complex protein class, which often has highly variable paratopes and lacks MSA data [45]. It has been demonstrated that by utilizing deep learning techniques on individual sequences, it is possible to circumvent the dependency on multiple sequence alignment-based methods. For instance, RGN2 [46] was developed as a single sequence approach which was trained using a language-based method and highlighted as a potential alternative to MSA. Furthermore, the results of this method were emphasized to be superior to those of AlphaFold2 and RoseTTAFold [47]. In addition, ESMFold [48] and OmegaFold [49] were designed subsequently using a similar approach. The Meta group created the Metagenomic Atlas which provides predicted structures for new metagenomic sequences. However, the use of such single sequence methods for protein prediction is often limited to smaller proteins (<400 amino acids) and often inadequate for orphan proteins, and Lin et al. are stated that

language models may have memorised multiple sequence alignments (MSA) to some extent [48].

5.2 Molecular Dynamics Trajectory-Based Approaches

Molecular dynamics trajectories are generated from molecular dynamics simulations. Trajectories depict atomic coordinates of a simulated molecular system at specific time intervals, represented as a series of sequential snapshots [50]. The trajectory information utilized encompasses the changes in the conformations of a protein, including unfolding or folding events, as the simulation progresses over time. The focus of the analysis is to determine if, and when, the protein reaches improved conformations from a range of starting model qualities [51]. If the issue of computationally expensive simulations can be resolved, then molecular dynamics trajectories will become practical for refinement purposes. However, solving these problems typically requires the integration of High Performance Computing systems with many (Graphical Processing Units) GPUs. Given that AF2 creates a single structure for a given protein sequence, using data sets for training that contain various dynamic conformational changes will be highly effective [52]. However, such time series data is not suitable for use in deep learning architectures, and some efforts are being made to provide them as inputs to deep learning methods [51].

5.3 Cryo-EM Based Approaches

With the advancements in cryo-EM density mapping and deep learning methods, there has been rapid progress in the field of structural biology, particularly in producing and refining protein structures at an atomic level with high accuracy [26,53]. as mentioned earlier. While AF2 often generates highly precise models using information from protein sequences, the main challenge with using AF2 is that the predicted structure is a single snapshot and may not match the conformation seen in a specific experimental context or conditions, such as those obtained through cryo-EM. To this end, improved structure reconstruction can be achieved by combining AlphaFold's predicted structure models with Cryo-EM data. A main example of this is the AlphaFold-Phenix collaboration, which has successfully improved the final predicted structure by using cryo-EM structures as templates [54]. However, determining atomic structures from cryo-EM maps with a resolution of 4

to 6 Å is a major challenge in the field of structural biology [55]. Therefore, the utilization of Cryo-EM data with these resolutions may pose difficulties. The Phenix version has been found to produce suitable results for density maps up to 4 Å, and the situation for structures beyond 4.5 Å is being investigated in the next Phenix version [25]. Another challenge in obtaining high-resolution 3D structures is the diversity of dynamic macromolecules. However, using single-particle cryo-EM data can reveal this heterogeneity [56].

5.4 Structure-Based Approaches

The realization of experimental level structure predictions in deep learning is achieved through not only protein language learning but also through the means of geometric learning [57]. Indeed, CNNs, have significantly improved the performance of methods addressing a range of problems, including sequence embedding for feature extraction and protein structure and function prediction. These networks overcome the limitations of traditional feature-based machine-learning methods by extracting task-specific features from protein sequences or structures. While various neural network applications, such as 3D CNNs, facilitate feature extraction, geometric-based neural networks can be a solution to this problem where they require a large amount of memory to explicitly process and store high-quality protein structures [58].

Despite the impressive success of CNNs in protein structure prediction, CNNs are not rotation-invariant and require data augmentation for the network to function properly. In contrast, graph neural networks (GNNs) perform convolution operations without relying on Cartesian coordinates by iteratively updating the properties of nodes and their neighbouring nodes. As a result, GNNs can handle graphs of any size and more naturally represent protein structures [59,60]. This can be a refinement of AF2 or non-AF2 models. AF2 structures are used as training examples for GNNs and are expanding the training sets for protein structure problems such as protein function prediction [61].

AF2 employed a neutralized refinement principle by utilizing a transformer that considers the translational and rotational symmetries of space. In contrast to NEMO, AF2 appears to refine

proteins using only a limited number of iterations, which may result in the displacement of protein coordinates in the non-realistic manner [12]. Most machine learning-based tools that improve current protein structures focus on improving the backbone structure without fully utilizing the structure of all atoms, including side-chains. There is still a gap in terms of improving the structure of all atoms. Refinement methods have recently been developed using equivariant networks in GNNs. These networks have been shown to improve AF2 structures according to various evaluation metrics, including all-atom contacts, bond length, atom clashes, torsion angles, and side-chain rotamers [18]. AF2 already implemented a new attention mechanism called “Invariant Point Attention (IPA)” for three-dimensional structures. This mechanism is similar to the SE(3) equivariant used in RoseTTAFold, but it is a simplified version [36]. When considering the potential for improved results in protein structure modelling and refinement. It can be possible that GNN together with structure-based information will be more widely utilized in the field of deep learning. Table 1 presents a list of tools that employ GNN architectures, specifically limited to the area of protein modelling.

6 CASP past and future

The Critical Assessment of Structure Prediction (CASP) evaluates the current state of protein structure modelling from amino acid sequences. The first CASP (CASP-1) competition was held in 1994, and the most recent CASP15 competition at the time of writing was held in 2022. It would be useful to emphasize once again the innovations that affected the field in the history of CASP. During the CASP12 period, the most significant advancements have been the utilization of residue contacts e.g., via RaptorX [62]. Subsequently in CASP13, the updated version of RaptorX and the first version of AF used predicted distance measurements for protein structure modelling [63], and AlQuraishi [12] also introduced end-to-end methods for protein modelling without using co-evolutionary information. In CASP14, the DeepMind team introduced their new AF2 architecture, which was also used in CASP-covid [64] and differs significantly from the original AF. This structure consists of two main modules, the Evoformer and structure modules, as well as performing recycling. In CASP14,

AF2 achieved outstanding performance [17], such that similar deep learning architectures were definitively established for protein structure prediction. After the DeepMind group shared the code source of AF2 with the community, almost all groups integrated the AF2 method into their own pipelines and were able to both benefit from and improve upon the models obtained. Following the success of AF2 in CASP14, the model refinement category was excluded from CASP15, nevertheless many groups continued to seek ways to further improve upon the AF2 models. There was still room for improvement to achieve experimental accuracy as deviations between predicted and experimental structures had been observed, in terms of backbone and/or side chain positions [24,65].

Despite DeepMind not participating in CASP15, AF2 has continued to have an impact and the groups with pipelines that integrated AF2 were among the most successful participants. Two primary methods were used by these groups in order to improve upon AF2: the utilization of more effective templates and/or MSA techniques, and the enhancement of AF2 through the implementation of dropout methods [66]. The option of using dropout during inference allows for activation of the stochastic component of the model, resulting in varying predictions. By iterating through different seeds, it is possible to sample different structural predictions from the uncertainty of the model or imprecision of the co-evolutionary constraints inferred from the input MSA [38]. Figure 3 shows some of the milestones that have influenced improvements in model quality in subsequent CASP competitions.

The use of the AF2 recycling process for improvement of models was another successful approach in CASP15 (Figure2). The term “recycling” refers to the process of tidying up some degree of disorganization before extracting valuable information. The initial recycling outputs, as well as the quality of templates/MSAs, play a crucial role in determining the amount of recycling required to successfully model the protein. High-quality templates, a large number of MSAs, or a well-known sequence may only require minor adjustments from an initial recycling output. However, missing templates or synthetic sequences may require extensive cleaning during recycling. By repeatedly

running the model through the network, recycling improves the model quality. In CASP15, our groups (McGuffin & MultiFOLD) used the AF2 recycling process to improve the quality of both tertiary and multimeric models beyond the quality of the baseline AF2 models [67,68]. Our MultiFOLD server [67] ranked as the 9th-best server group on monomeric targets according to the GDT_TS scores and the 8th-best server group for multimers by the assessors' formula, outperforming both the baseline NBIS-AF2- standard (AlphaFold2) group and NBIS-AF2-multimer (AlphaFold2- Multimer) group across the board. Therefore, the MultiFOLD server has been independently verified to produce higher quality tertiary and quaternary structure models than AlphaFold2 according to all measures [67].

7. Conclusion

The development of methods for protein structure prediction has been ongoing for several decades, with continual improvements that have recently resulted in the availability of models for tertiary structures with near experimental quality for many targets. The efficacy of deep learning methods in this field was demonstrated with the success of AF2 in CASP14, and after the source code was made publicly available by DeepMind, the application of AF2 and other deep learning-based methods has become widespread. Subsequently, a process of improving upon AF2 began, with various types of data utilized to obtain more accurate 3D models, with the most noteworthy example being the use of Cryo-EM data. In CASP15, more studies were conducted on co-evolution information from MSA and geometric representation of structures. The methods used included generative models for obtaining more homologous sequences, protein language modelling, and graph neural network variations, as well as recycling and dropout methods for improving or refining structures. Although the validity of using more traditional refinement methods, such as MD, has not been fully proven for deep learning-based models, they are now only commonly used as minimization methods, due to the inadequacy of force fields and high computational costs required. The recent improvements made to AF2 and other deep learning-based methods will be crucial in fields such as drug design and will continue to be relevant as we move towards experimental quality prediction of quaternary

structures.

Acknowledgments

This work was supported by the Republic of Turkey Ministry of National Education (to A.G.G.) and the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/T018496/1 to L.J.M.].

References

1. Kuhlman B, Bradley P (2019) Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology 20 (11):681-697. doi:10.1038/s41580-019-0163-x

2. Chou P, Fasman GD (2009) Amino acid sequence. Adv Enzymol Relat Areas Mol Biol 47:45

3. Dunker AK, Lawson JD, Brown CJ et al. (2001) Intrinsically disordered protein. J Mol Graph Model 19 (1):26-59. doi:10.1016/s1093-3263(00)00138-8

4. Dunker AK, Oldfield CJ (2015) Back to the Future: Nuclear Magnetic Resonance and Bioinformatics Studies on Intrinsically Disordered Proteins. Adv Exp Med Biol 870:1-34. doi:10.1007/978-3-319-20164-1_1

5. Bondos SE, Dunker AK, Uversky VN (2021) On the roles of intrinsically disordered proteins and regions in cell communication and signaling. Cell Communication and Signaling 19 (1):88. doi:10.1186/s12964-021-00774-3

6. Coskuner O, Uversky VN (2019) Intrinsically disordered proteins in various hypotheses on the pathogenesis of Alzheimer's and Parkinson's diseases. Prog Mol Biol Transl Sci 166:145-223. doi:10.1016/bs.pmbts.2019.05.007

7. Basile W, Salvatore M, Bassot C et al. (2019) Why do eukaryotic proteins contain more intrinsically disordered regions? PLOS Computational Biology 15 (7):e1007186. doi:10.1371/journal.pcbi.1007186

8. Wooley JC, Ye Y (2007) A Historical Perspective and Overview of Protein Structure Prediction. In: Xu Y, Xu D, Liang J (eds) Computational Methods for Protein Structure Prediction and Modeling: Volume 1: Basic Characterization. Springer New York, New York, NY, pp 1-43. doi:10.1007/978-0-387-68372-0_1

9. Sanger F, Thompson EO, Kitai R (1955) The amide groups of insulin. Biochem J 59 (3):509-518. doi:10.1042/bj0590509

10. Anfinsen CB, Redfield RR, Choate WL et al. (1954) Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. J Biol Chem 207 (1):201-210

11. Pearce R, Zhang Y (2021) Deep learning techniques have significantly impacted protein structure prediction and protein design. Curr Opin Struct Biol 68:194-207. doi:10.1016/j.sbi.2021.01.007

12. AlQuraishi M (2019) End-to-End Differentiable Learning of Protein Structure. Cell Systems 8 (4):292-301.e293. doi:https://doi.org/10.1016/j.cels.2019.03.006

13. Xu J (2019) Distance-based protein folding powered by deep learning. Proceedings of the National Academy of Sciences 116 (34):16856-16865. doi:doi:10.1073/pnas.1821309116

14. Senior AW, Evans R, Jumper J et al. (2020) Improved protein structure prediction using potentials from deep learning. Nature 577 (7792):706-710. doi:10.1038/s41586-019-1923-7

15. Gomes P, Gomes DEB, Bernardi RC (2022) Protein structure prediction in the era of AI: Challenges and limitations when applying to in silico force spectroscopy. Front Bioinform 2:983306. doi:10.3389/fbinf.2022.983306

16. Bhattacharya D (2019) refineD: improved protein structure refinement using machine learning based restrained relaxation. Bioinformatics 35 (18):3320-3328. doi:10.1093/bioinformatics/btz101

17. Jumper J, Evans R, Pritzel A et al. (2021) Applying and improving AlphaFold at CASP14. Proteins 89 (12):1711-1721. doi:10.1002/prot.26257

18. Wu T, Guo Z, Cheng J (2023) Atomic protein structure refinement using all-atom graph representations and SE(3)-equivariant graph transformer. Bioinformatics 39 (5). doi:10.1093/bioinformatics/btad298

19. Adiyaman R, McGuffin LJ (2021) ReFOLD3: refinement of 3D protein models with gradual restraints based on predicted local quality and residue contacts. Nucleic Acids Research 49 (W1):W589-W596. doi:10.1093/nar/gkab300

20. Feig M, Mirjalili V (2016) Protein structure refinement via molecular-dynamics simulations: What works and what does not? Proteins 84 Suppl 1 (Suppl 1):282-292. doi:10.1002/prot.24871

21. Townshend RJL, Bedi R, Suriana PA et al. (2019) End-to-end learning on 3D protein structure for interface prediction. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., p Article 1401

22. Mirabello C, Wallner B (2019) rawMSA: End-to-end Deep Learning using raw Multiple Sequence Alignments. PLOS ONE 14 (8):e0220182. doi:10.1371/journal.pone.0220182

23. Ingraham J, Riesselman AJ, Sander C et al. Learning Protein Structure with a Differentiable Simulator. In: International Conference on Learning Representations, 2018.

24. Shuvo MH, Gulfam M, Bhattacharya D (2021) DeepRefiner: high-accuracy protein structure refinement by deep network calibration. Nucleic Acids Research 49 (W1):W147-W152. doi:10.1093/nar/gkab361

25. Terwilliger TC, Poon BK, Afonine PV et al. (2022) Improved AlphaFold modeling with implicit experimental information. Nature Methods 19 (11):1376-1382. doi:10.1038/s41592-022-01645-6

26. Zhang B, Liu D, Zhang Y et al. (2022) Accurate flexible refinement for atomic-level protein structure using cryo-EM density maps and deep learning. Briefings in Bioinformatics 23 (2). doi:10.1093/bib/bbac026

27. Jumper J, Evans R, Pritzel A et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873):583-589. doi:10.1038/s41586-021-03819-2

28. Xu G, Wang Q, Ma J (2021) OPUS-Rota4: a gradient-based protein side-chain modeling framework assisted by deep learning-based predictors. Briefings in Bioinformatics 23 (1). doi:10.1093/bib/bbab529

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

30. Morehead A, Chen X, Wu T et al. (2022) EGR: Equivariant Graph Refinement and Assessment of 3D Protein Complex Structures. doi:10.48550/arXiv.2205.10390

31. Bryant P, Pozzati G, Zhu W et al. (2022) Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. Nature Communications 13 (1):6028. doi:10.1038/s41467-022-33729-4

32. Adiyaman R, McGuffin LJ (2019) Methods for the Refinement of Protein Structure 3D Models. Int J Mol Sci 20 (9). doi:10.3390/ijms20092301

33. Cárdenas R, Martínez-Seoane J, Amero C (2020) Combining Experimental Data and Computational Methods for the Non-Computer Specialist. Molecules 25 (20). doi:10.3390/molecules25204783

34. Vetrivel I, Mahajan S, Tyagi M et al. (2017) Knowledge-based prediction of protein backbone conformation using a structural alphabet. PLoS One 12 (11):e0186215. doi:10.1371/journal.pone.0186215

35. Bepler T, Berger B (2021) Learning the protein language: Evolution, structure, and function. Cell Systems 12 (6):654-669.e653. doi:https://doi.org/10.1016/j.cels.2021.05.017

36. Rubiera CO (2021) AI3SD Video: How good are protein structure prediction methods at predicting folding pathways? Paper presented at the AI 4 Proteins Seminar Series 2021, 14/04/21 - 17/06/21

37. Mirdita M, Steinegger M, Söding J (2019) MMseqs2 desktop and local web server app for fast, interactive sequence searches. Bioinformatics 35 (16):2856-2858. doi:10.1093/bioinformatics/bty1057

38. Mirdita M, Schütze K, Moriwaki Y et al. (2022) ColabFold: making protein folding accessible to all. Nature Methods 19 (6):679-682. doi:10.1038/s41592-022-01488-1

39. Wu Z, Johnston KE, Arnold FH et al. (2021) Protein sequence design with deep generative models. Current Opinion in Chemical Biology 65:18-27. doi:https://doi.org/10.1016/j.cbpa.2021.04.004

40. Consortium TU (2022) UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research 51 (D1):D523-D531. doi:10.1093/nar/gkac1052

41. Hou Q, Pucci F, Pan F et al. (2022) Using metagenomic data to boost protein structure prediction and discovery. Computational and Structural Biotechnology Journal 20:434-442. doi:https://doi.org/10.1016/j.csbj.2021.12.030

42. Yang P, Zheng W, Ning K et al. (2021) Decoding the link of microbiome niches with homologous sequences enables accurately targeted protein structure prediction. Proceedings of the National Academy of Sciences 118 (49):e2110828118. doi:doi:10.1073/pnas.2110828118

43. Bryant P, Pozzati G, Elofsson A (2022) Improved prediction of protein-protein interactions using AlphaFold2. Nature Communications 13 (1):1265. doi:10.1038/s41467-022-28865-w

44. Basu S, Wallner B (2016) DockQ: A Quality Measure for Protein-Protein Docking Models. PLOS ONE 11 (8):e0161879. doi:10.1371/journal.pone.0161879

45. Jin W, Barzilay DR, Jaakkola T (2022) Antibody-Antigen Docking and Design via Hierarchical Structure Refinement. Paper presented at the Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research

46. Chowdhury R, Bouatta N, Biswas S et al. (2022) Single-sequence protein structure prediction using a language model and deep learning. Nature Biotechnology 40 (11):1617-1623. doi:10.1038/s41587-022-01432-w

47. Baek M, DiMaio F, Anishchenko I et al. (2021) Accurate prediction of protein structures and interactions using a three-track neural network. Science 373 (6557):871-876. doi:10.1126/science.abj8754

48. Lin Z, Akin H, Rao R et al. (2023) Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 379 (6637):1123-1130. doi:doi:10.1126/science.ade2574

49. Wu R, Ding F, Wang R et al. (2022) High-resolution *evol* structure prediction from primary sequence. bioRxiv:2022.2007.2021.500999. doi:10.1101/2022.07.21.500999

50. Likhachev IV, Balabaev NK, Galzitskaya OV (2016) Available Instruments for Analyzing Molecular Dynamics Trajectories. Open Biochem J 10:1-11. doi:10.2174/1874091x01610010001

51. Pfeiffenberger E, Bates PA (2018) Predicting improved protein conformations with a temporal deep recurrent neural network. PLOS ONE 13 (9):e0202652. doi:10.1371/journal.pone.0202652

52. Sathvik Kolli AL, Xinyang Geng, Aviral Kumar, Sergey Levine (2022) Data-Driven Optimization for Protein Design: Workflows, Algorithms and Metrics. Paper presented at the ICLR Workshop on Machine Learning for Drug Discovery

53. Glaeser RM (2016) How good can cryo-EM become? Nature Methods 13 (1):28-32. doi:10.1038/nmeth.3695

54. Giri N, Roy RS, Cheng J (2023) Deep learning for reconstructing protein structures from cryo-EM density maps: Recent advances and future directions. Current Opinion in Structural Biology 79:102536. doi:https://doi.org/10.1016/j.sbi.2023.102536

55. Alshammari M, He J, Wriggers W Refinement of AlphaFold2 Models against Experimental Cryo-EM Density Maps at 4-6Å Resolution. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 6-8 Dec. 2022 2022. pp 3423-3430. doi:10.1109/BIBM55620.2022.9995676

56. Doerr A (2016) Single-particle cryo-electron microscopy. Nature Methods 13 (1):23-23. doi:10.1038/nmeth.3700

57. Laine E, Eismann S, Elofsson A et al. (2021) Protein sequence-to-structure learning: Is this the end(-to-end revolution)? Proteins: Structure, Function, and Bioinformatics 89 (12):1770-1786. doi:https://doi.org/10.1002/prot.26235

58. Gligorijević V, Renfrew PD, Kosciolek T et al. (2021) Structure-based protein function prediction using graph convolutional networks. Nature Communications 12 (1):3168. doi:10.1038/s41467-021-23303-9

59. Réau M, Renaud N, Xue LC et al. (2022) DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces. Bioinformatics 39 (1). doi:10.1093/bioinformatics/btac759

60. Wang X, Flannery ST, Kihara D (2021) Protein Docking Model Evaluation by Graph Neural Networks. Frontiers in Molecular Biosciences 8. doi:10.3389/fmolb.2021.647915

61. Ma W, Zhang S, Li Z et al. (2022) Enhancing Protein Function Prediction Performance by Utilizing AlphaFold-Predicted Protein Structures. Journal of Chemical Information and Modeling 62 (17):4008-4017. doi:10.1021/acs.jcim.2c00885

62. Källberg M, Wang H, Wang S et al. (2012) Template-based protein structure modeling using the RaptorX web server. Nature Protocols 7 (8):1511-1522. doi:10.1038/nprot.2012.085

63. Pakhrin SC, Shrestha B, Adhikari B et al. (2021) Deep Learning-Based Advances in Protein Structure Prediction. Int J Mol Sci 22 (11). doi:10.3390/ijms22115553

64. Kryshtafovych A, Moult J, Billings WM et al. (2021) Modeling SARS-CoV-2 proteins in the CASP-commons experiment. Proteins 89 (12):1987-1996. doi:10.1002/prot.26231

65. Schreiner M (2022) CASP15: AlphaFold's success spurs new challenges in protein-structure prediction. https://the-decoder.com/casp15-alphafolds-success-brings-new-challenges/, vol 2023.

66. Elofsson A (2022) Protein Structure Prediction until CASP15. arXiv:221207702. doi:https://doi.org/10.48550/arXiv.2212.07702

67. McGuffin LJ, Edmunds NS, Genc AG et al. (2023) Prediction of protein structures, functions and interactions using the IntFOLD7, MultiFOLD and ModFOLDdock servers. Nucleic Acids Research 51 (W1):W274-W280. doi:10.1093/nar/gkad297

68. Adiyaman R, Edmunds NS, Genc AG et al. (2023) Improvement of protein tertiary and quaternary structure predictions using the ReFOLD refinement method and the AlphaFold2 recycling process. Bioinformatics Advances 3 (1). doi:10.1093/bioadv/vbad078

69. Jing X, Xu J (2021) Fast and effective protein model refinement using deep graph neural networks. Nature Computational Science 1 (7):462-469. doi:10.1038/s43588-021-00098-9

70. Johansson-Åkhe I, Wallner B (2022) InterPepScore: a deep learning score for improving the FlexPepDock refinement protocol. Bioinformatics 38 (12):3209-3215. doi:10.1093/bioinformatics/btac325

71. Chinery L, Wahome N, Moal I et al. (2022) Paragraph—antibody paratope prediction using graph neural networks with minimal feature vectors. Bioinformatics 39 (1). doi:10.1093/bioinformatics/btac732

72. Igashov I, Olechnovič K, Kadukova M et al. (2021) VoroCNN: deep convolutional neural network built on 3D Voronoi tessellation of protein structures. Bioinformatics 37 (16):2332-2339. doi:10.1093/bioinformatics/btab118

73. Sunny S, Prakash PB, Gopakumar G et al. (2023) DeepBindPPI: Protein–Protein Binding Site Prediction Using Attention Based Graph Convolutional Network. The Protein Journal. doi:10.1007/s10930-023-10121-9

Table1 A list of tools that employ graph neural networks for the prediction of specific protein modelling, protein structure refinement, or related intermediate processes.

GNN-Based Tool	Brief Method	Reference	Link
RoseTTAfold	SE(3)-equivariant graph transformer to improve backbone structure without leveraging side-chain	[47]	https://github.com/RosettaCommons/RoseTTAFold
GNNRefine	Its main aim is to provide per residue accuracy score and distance error between pair residues in order via Convolutional Neural Network to guide the refinement process of Rosetta.	[69]	http://raptorx.uchicago.edu
ATOMRefine	It employed the combination of a transformer with 3D- equivariant (not dependent on translation and rotation) by learning main structural information in order to output refined coordinates.	[18]	https://github.com/BioinfoMachineLearning/ATOMRefine
DeepRank-GNN	The main point is to convert interface protein-protein complexes from 3D coordinates PDB file into graph via GNN	[59]	https://github.com/DeepRank/deeprank
InterPepScore	It generates DockQ score via GNN whose training set was trajectories from initial coarse by perturbing the native peptide (the improvement of FlexPepDock refinement tool by adding additional scoring term).	[70]	http://wallnerlab.org/InterPepScore
Equivariant Graph Refiner (EGR)	The integration of SE(3)-equivariance into message-passing neural networks and using refinement loss between the model's coordinates and ground truth coordinates for the refinement of 3-D structures.	[30]	https://github.com/BioinfoMachineLearning/DeepRefine
Paragraph	Main aim is to predict paratope. It uses equivariant graph neural network to predict the possible relevant residues of paratope when the antibody structure is taken as input	[71]	www.github.com/oxpig/Paragraph
VoroCNN	It predicts the local-CAD score for all residues in the original structure using a graph neural network, even if the main part consists of Convolutional Neural network.	[72]	https://team.inria.fr/nano-d/software/vorocnn/

DeepBindPPI	It includes the inclusion of an attention mechanism to GCN to predict the binding sites to proteins.	[73]	https://github.com/Sharon1989Sunny/DBP

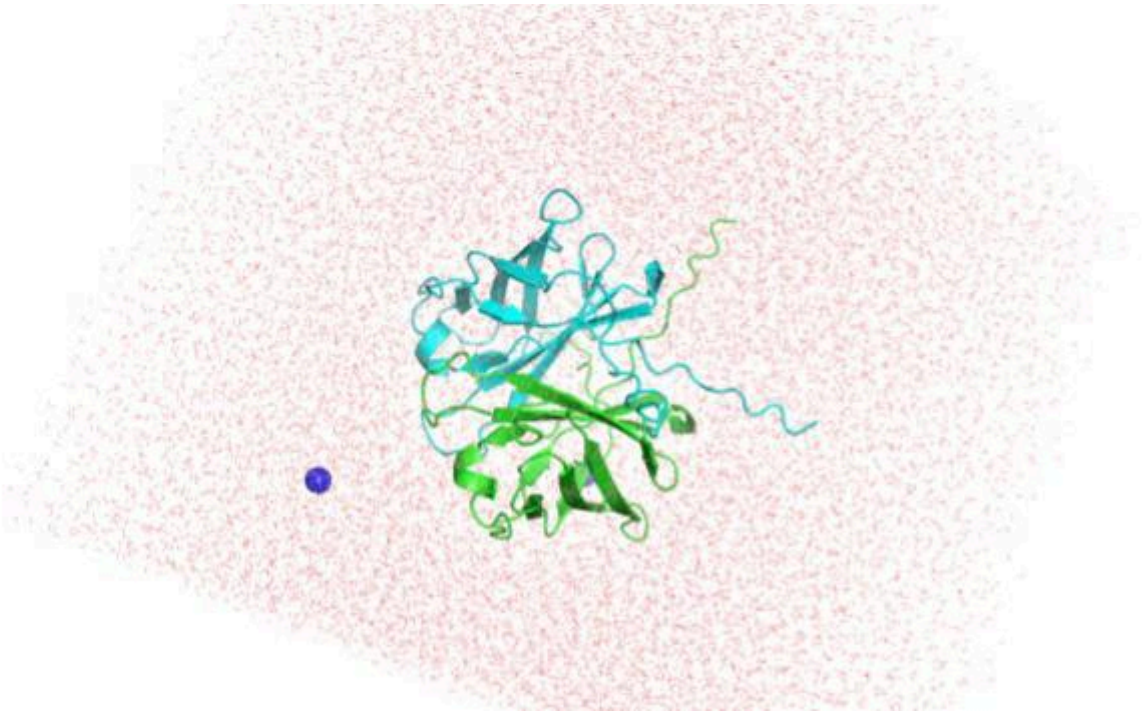

Fig. 1 View of Molecular Dynamic Simulation which is the traditional most popular refinement method. It depicts the multimeric protein structure (T1078) and its environment resulting from molecular dynamics simulations. It is a visual representation of the simulation output from Pymol. In this simulation, the 3D protein structure was generated within a cubic framework, surrounded by water molecules (red colour) and ions (blue spherical structures) according to chosen parameters, which corresponds to a single MD trajectory for the given protein.

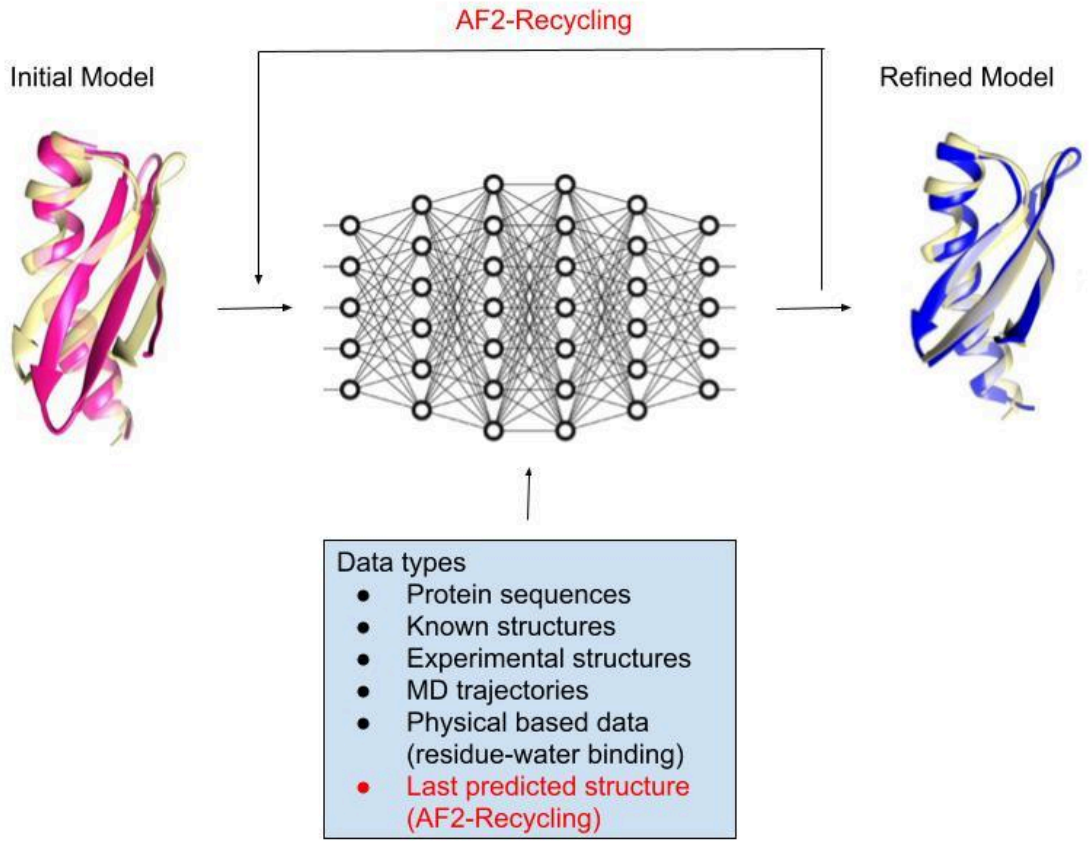

Fig. 2 The figure illustrates the general workflow of methods for improving a modelled structure using deep learning techniques. The aim is to obtain a protein structure of higher quality using deep neural networks designed by the researcher. The use of deep neural networks for various types of data, where their efficient utilization for training has been researched, has become widespread after AF2. The red-highlighted component represents the iterative improvement process referred to as recycling, which is included in the AF2 algorithm.

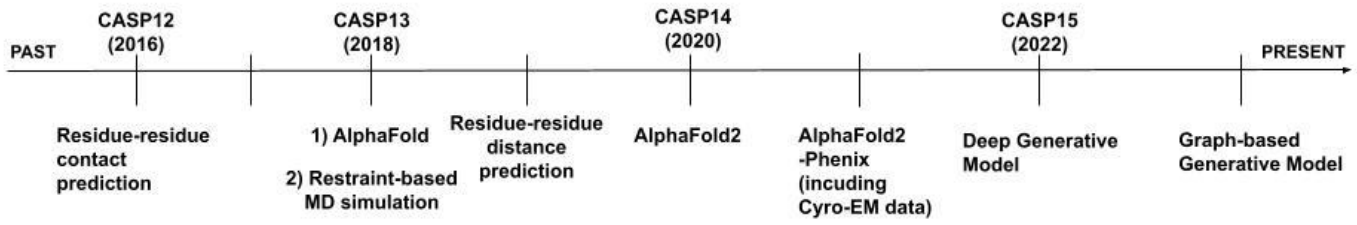

Fig. 3 This graph encompasses the significant milestones and methods, either directly or indirectly impacting structure improvement, in the field of protein bioinformatics over time for structure refinement. The critical turning points have been selected as the CASP competitions and the methods and tools that have been commonly effective in these competitions, as the competition has gained substantial importance in guiding the community.