

## **A glance into the evolution of template-free protein structure prediction methodologies**

Surbhi Dhingra, Ramanathan Sowdhamini, Frédéric Cadet, Bernard Offmann

► **To cite this version:**

Surbhi Dhingra, Ramanathan Sowdhamini, Frédéric Cadet, Bernard Offmann. A glance into the evolution of template-free protein structure prediction methodologies. *Biochimie*, Elsevier, 2020, 175, pp.85-92. 10.1016/j.biochi.2020.04.026 . hal-03142832

**HAL Id: hal-03142832**

**<http://hal.univ-nantes.fr/hal-03142832>**

Submitted on 16 Feb 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# A glance into the evolution of template-free protein structure prediction methodologies

Surbhi Dhingra<sup>1</sup>, Ramanathan Sowdhamini<sup>2</sup>, Frédéric Cadet<sup>3,4,5</sup>, and Bernard Offmann\*<sup>1</sup>

<sup>1</sup>Université de Nantes, CNRS, UFIP, UMR6286, F-44000 Nantes, France

<sup>2</sup>Computational Approaches to Protein Science (CAPS), National Centre for Biological Sciences (NCBS), Tata Institute for Fundamental Research (TIFR), Bangalore 560-065, India

<sup>3</sup>University of Paris, BIGR—Biologie Intégrée du Globule Rouge, Inserm, UMR\_S1134, Paris F-75015, France

<sup>4</sup>Laboratory of Excellence GR-Ex, Boulevard du Montparnasse, Paris F-75015, France

<sup>5</sup>DSIMB, UMR\_S1134, BIGR, Inserm, Faculty of Sciences and Technology, University of La Reunion, Saint-Denis F-97715, France

## Abstract

Prediction of protein structures using computational approaches has been explored for over two decades, paving a way for more focused research and development of algorithms in comparative modelling, *ab initio* modelling and structure refinement protocols. A tremendous success has been witnessed in template-based modelling protocols, whereas strategies that involve template-free modelling still lag behind, specifically for larger proteins (> 150 a.a.). Various improvements have been observed in *ab initio* protein structure prediction methodologies over-time, with recent ones attributed to the usage of deep learning approaches to construct protein backbone structure from its amino acid sequence. This review highlights the major strategies undertaken for template-free modelling of protein structures while discussing few tools developed under each strategy. It will also briefly comment on the progress observed in the field of *ab initio* modelling of proteins over the course of time as seen through the evolution of CASP platform.

This paper is dedicated to the memory of Anna Tramontano (1957-2017) who was an Italian computational biologist and chair professor of biochemistry at the Sapienza University of Rome.

Declarations of interest: none

Keywords: protein structure prediction; ab-initio; template-free modelling; artificial intelligence

---

\*corresponding author : [bernard.offmann@univ-nantes.fr](mailto:bernard.offmann@univ-nantes.fr)

Abbreviations: Critical Assessment of protein Structure Prediction (CASP), Template-based modelling (TBM), Template-free modelling (TBM), Fragment-based approaches (FBA), Artificial Intelligence (AI).

## 1 Introduction

Proteins are complex biomolecules that play a crucial role in building, strengthening, maintaining, protecting and repairing a living entity. Each protein folds into a specific three-dimensional structure owing to its amino acid composition. This in turn corresponds to a specific function, collectively termed as sequence-structure-function paradigm [1]. The relationship between protein sequence and its corresponding secondary and tertiary structure is termed as second genetic code [2]. A major gap exists in our knowledge of the science behind protein folding based on its sequence. Research focused in deciphering the second genetic code has been budding for past few decades by means of various schemes.

Advent of genomics has led to the availability of large deposit of sequence data online. This helps in easy classification of proteins and in approximating their functional annotation. A considerable amount of this classification is based on shared sequence similarity (and conserved domain search) between two or more sequences. Currently, UniProtKB/TrEMBL database is enriched with around 170 million sequence data [3]. Yet protein functionality remains unclear primarily due to the lack of structural description at the atomic levels. The equivalent structural database, RCSB [4] (<https://www.rcsb.org>) documents around 160,000 structures to date belonging to well defined protein families. There is also an ever increasing gap between protein sequence and structure data availability due to considerable growth observed in sequencing techniques.

Scientific community has always relied on experimental approaches to deliver high resolution protein structures. Structural data deposited in data banks are only accountable when verified through experiments like X-Ray [5], NMR [6] etc. Time and again these techniques have been proven to be most efficient in getting relevant spatial characterisation of a protein. On the other hand, they also have remained stagnant in terms of improvements due to being heavily restricted by time and manpower requirements [7]. A recent introduction of Cryo-EM has fostered an acceleration of protein structure determination process [8]. The core of this technique lies in photographing frozen molecules to determine their structure. Nonetheless, the approach is relatively new and usually generates lower resolution structures than those benchmarked by other experimental techniques.

Twentieth century has witnessed a blooming era for scientific community indulging in computational approaches for approximation of protein structures. Anfinsen in 1972 laid the foundation for protein structure prediction by correctly refolding ribonuclease molecule from its sequence [9]. As stated in the paper, “the native conformation is determined by the totality of inter-atomic interactions and hence by the amino acid sequence in a given environment” [10]. In other words, a protein attains its conformational nativity when its environment is at its lowest Gibbs free energy levels. Another statement put forward in their work was that a protein structure is only stable and functional in the environment it was chosen during natural selection. Despite knowing the physical environment requirement for folding a protein sequence, it remains a challenge to fold them into their functional form. Therefore, limiting the understanding of the sequence-structure-function paradigm [11, 12].

Computational approaches for protein structure prediction can broadly be categorized into two groups: Template-Based Modelling (TBM) [13, 14] and Template Free Modelling/Free-Modelling (FM) [15]. A representative flowchart of the categorization is illustrated in Figure 1. This classification has been adopted by well-known biennial competition of protein structure prediction, Critical Assessment of protein Structure Prediction (CASP) [16] [17–20]. Results

from this competition benchmark the improvement in the field of computational protein structure prediction [19,21]. Majority of progress witnessed in this field is in construction of protein models using templates sharing high sequence similarities with unknown protein. The basis behind the approach is that similar sequence tend to fold in a similar manner. This tendency of proteins to envelop into similar folds reduces with shared sequence similarity, though there exist cases of proteins having same folds even when their shared sequence similarity is low.

TBM, as the name suggests, makes use of template to predict 3D models. Single or multiple homologous protein sharing high sequence similarity are aligned to the unknown protein sequence predicting likely models [13]. Structures predicted through TBM usually have a good resolution and might fall into same functional classes. But there is little progress made when it comes to predicting new protein folds or structures. TBM is an effective approach as long as the query shares at-least 30% sequence identity with the template [22]. On the basis of shared sequence identity, it can be classified into Homology Modelling (HM) [23–25], Comparative Modelling (CM) [26, 27] and Threading approaches (fold-recognition) [28–31]. Each sub-class follows similar methodology into prediction of protein three-dimensional organisation from its primary one-dimensional sequence. One might argue that HM and CM are two terms for one and the same approach. It is true to a great extent except that homology modelling is defined when template shares an ancestry with the query being modelled whereas in case of CM, the query sequence has no identified evolutionary relationship with the template but only shared sequence similarity. So far, comparative modelling has been the most successful computational protein structure prediction approach available [22]. The third category of TBM is fold-recognition/threading which follows the idea of picking template structures based on their fit with the protein sequence in question. It is basically a comparison of 1D protein sequence to 3D template structure.

## 2 *Ab initio* Protein Structure Prediction

A significant amount of sequence data does not share homology with well-studied protein families. This called for development of approaches which could help predicting protein structures with minimal or no known information. Such approaches fall into the second major class of computational protein structure prediction called “Template-Free modelling/Free-modelling” (TFM/FM). The word “free” used in the name indicates the initial take on such algorithms to rely on physical laws to determine protein structures. Though, most of the algorithms developed around it are guided by structural information. In this review we will touch into the evolution of Free-modelling and the approaches that have been used to predict 3D models. Throughout this review Free-modelling, *ab initio* modelling and *de novo* modelling will be used interchangeably to discuss template-free modelling approaches.

Template-free modelling comprises of algorithms/pipeline/methods for generating protein models with no known structural homologs available. Mainly these approaches focused on using physics based principles and energy terms to model proteins. The nomenclature remains debatable as in several cases, information from known structures is used in one way or another. This review is considering the following definition as best suited to describe our understanding of TFM: “*Ab initio* protein structure prediction or Free modelling (FM) can most appropriately be defined as an effort to construct 3D structure without using homologous proteins as template” [14, 22, 32–34]. FM approaches majorly depend on designing algorithms with ability to rapidly locate global energy minimum and a scoring function capable of selecting best available conformation from the several generated models [35–37].

The aim of free-modelling protocols is to predict the most stable protein spatial arrangement with lowest free energy. The major challenge faced while developing *ab initio* approaches is searching conformational space which is usually huge considering the dynamic nature of

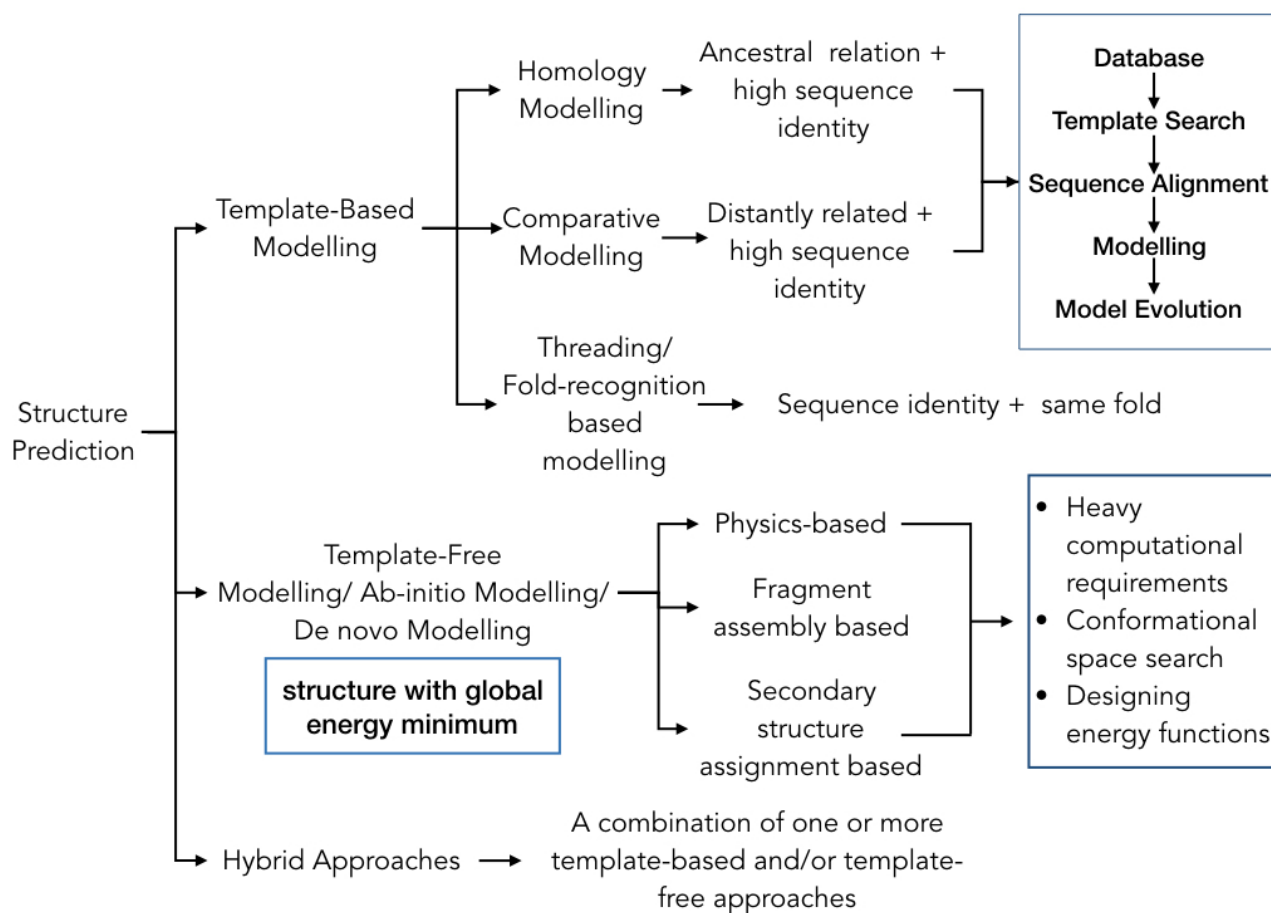


Figure 1: Computation Protein Structure Prediction Approaches.

This figure provides a broad classification of few computational protein structure prediction strategies developed and used to determine protein structure. The two major classifications of these methodologies lie under the domain of Template-Based Modelling and Template-free modelling. Each of these categories can further be split into a set of strategies based on the basic principle followed by the parent approaches for structure prediction.

proteins. Since, these approaches involve building the protein structure from scratch, focus is laid on building effective energy functions to minimise conformational search space and facilitate accurate folding [22, 34]. *Ab initio* algorithms can also be influenced from experimental data available in the form of abstract NMR restraints, predicted residue-residue contact maps, Cryo-EM density maps etc. [38–40].

## 2.1 Strategies for *Ab initio* Prediction of Protein Structure

Free modelling has witnessed a major bloom in the past era owing to several strategies developed for structure prediction, few of them have been stated in the Table 1. Initially the scientific community resorted to use pure physics based laws, MD simulations etc. to explore the atomic dynamics of protein molecules. The prediction horizon expanded with time into utilizing restraints like  $C\alpha$ - $C\alpha$  distance, dihedral angles, solvent interactions, side-chain atoms, contact map information and more from available structures. The newer fundamentals involved building saturated library of structural information in the form of small fragments, secondary structural elements, motifs, foldons etc. Below we have broadly classified the *ab initio* protein

Table 1: Strategies available for protein structure prediction. First two columns provide the name and a brief description of few algorithms developed overtime for *ab initio* protein structure prediction. Third column states the length which for most tools is indicative state-of-the-art results observed throughout the field, i.e, able to achieve a GDT\_TS score of 30% and above. For most part high accuracy models generation is still limited to small protein (< 150 residue long).

Algorithms/ Servers	Strategy/ Approach	Maximum predicted length	Reference
Rosetta	Fragment-assembly using MC simulations, all-atom energy function to determine structure and clustering of models	~ 250	[41-48]
Quark	Fragment-based assembly using REMC simulation guided by knowledge based potentials	upto 150	[7, 49]
EdaRose	Utilising EdaFold algorithm for fragment based assembly with cluster based and energy based variations	upto 150	[20]
Chunk-Taser	Hybrid approach using restraints derived from super secondary structure chunks as well as by threading the templates	< 200	[50]
UNRES	Physics-based conformational space search using UNRES energy function	< 80	[51]
BCL::FOLD	Assembling secondary structural elements using MC sampling and knowledge-based energy functions	~ 200	[15, 40, 52]
SS-Thread	Prediction of contacting pairs of $\alpha$ -helix and $\beta$ -strand	< 200	[53]
UniCon3D	Using foldons and probabilistic models to capture local backbone structural preference and side chain conformation search space	~ 200	[34]
Bhageerath-H	Hybrid approach involving combination of several tools developed in the lab with the goal to reduce conformational search space	~ 350	[37]
Smotiff	Fragment-Based approach developed using saturated library of super secondary structure fragments	~ 150	[54]
Touchstone	Secondary and tertiary restraints prediction through threading-based approaches	~ 150	[55, 56]
PconsFold	Evolutionary based structure prediction pipeline using PconsC contact predictions on Rosetta folding protocols	~ 250	[57]
EdaFold	Fragment-based all atom energy function to produce atomic models	upto 150	[58, 59]
Astro-Fold	folding amino acid sequences using first-principle based approaches	~ 150	[60, 61]
DMFold	Deep learning based protein modelling by incorporating predicted structural constraints like inter-atomic distance bonds, hydrogen bonds and torsion angles	upto 200	[62]

structure prediction approaches based on the core methodologies used to develop them.

## 2.2 Physics Based Approaches

These formed the basis of initial algorithms built under the emerging field. The main idea behind developing these physics-based approaches is to rely on MD simulations to trace the folding path of the proteins. The philosophy backing their design is to obtain lowest energy conformation model by folding the protein sequence using quantum mechanism and coulomb potential [63,64]. But due to high computational requirements, the field majorly relies on inter atomic interactions and force fields to solve the protein folding problem.

Free energy calculations have been explored from the very beginning of computational protein structure prediction evolution. It is believed that these approaches can go beyond documented structures and capture novel folds and patterns by exploring the inherent dynamic motion of proteins [65,66]. Despite the availability of better computing, physics based approach continues to lag behind due to the amount of time required to reach the native state along-with the meddling of erroneous force-field that restrict the model to attain it [12,67–69].

MELD (Modelling Employing Limited Data) [65] is a recently developed physics-based protein structure prediction approach which uses Bayesian law to tap into atomic molecular dynamics of proteins for structural modelling. It has proven to be effective in determining high resolution structures of proteins up to 260 residues long [65]. Similar effort was made by David Shaw’s group where they utilised different sets of restraints to reduce the MD simulation runs and prevent the model from getting trapped in non-native energy state [70]. H. Nguyen et al demonstrated that the combination of an implicit solvent and a force field can result in near-native models in-case of small proteins (less than 100 amino acids) [71]. Another group showed that simulation time can be reduced and energy landscapes can be managed using residue-specific force field (RSFF1) in explicit solvent and Replica exchange molecular dynamics (REMD) [72].

## 2.3 Fragment Based Approaches (FBA)

It is by far the most successful strategy used for template-free prediction of protein structure. This approach revolves around the construction of fragment libraries of varied lengths, where each fragment represents a pseudo-structure. The idea is to map information from protein fragments instead of using entire templates for constructing protein model. Segments of query sequences are replaced by the fragment coordinates recorded in the fragment library or by its predicted fold. Since, it is computationally exhaustive to go through all possible protein fold conformations for a structure built from scratch, fragmenting the sequence limits the number of folding patterns thus reducing the computational expense. Bowie and Eisenberg introduced Fragment-Based assembly approach to predict protein structures [41]. They used fragments of length 9 to 25 from a database of known proteins and an energy function (composed of 6 terms) that can guide building of energetically stable models [41]. This attempt set path for the evolution of computational 3D-modelling of protein structures using fragments.

Through the years several fragment-based approaches have been developed; few of which have done exceedingly well and remain the best options for *ab initio* protein structure prediction to date. The basic idea behind these algorithms remains the same and typically varies with fragment type, length and scoring functions used to generate energetically minimised stable structure. Rosetta [43,44], one of the most renowned fragment based approach, uses fragment libraries of length 3 and 9. It follows a Monte Carlo simulation based strategy to predict globally minimised protein models. The scoring function used in Rosetta is based on Bayesian separation of total energy into individual components. Its highest achievement has been noted during

CASP11, where it correctly predicted the structure of a 256 amino acid long sequence [73].

SmotifsTF [54] produces library of supersecondary structure fragments known as Smotifs to built probable models. The fragment library construction and utilisation is based on fragment assembly protocols. The fragment collection is governed by weak sequence similarities generating fragments on average of 25 residues in length. QUARK [49] has more dynamic fragment length range of up to 20 residues which are assembled using replica-exchange Monte Carlo simulations guided by knowledge-based force-field. It has also been ranked as the best predictor in FM category for both CASP9 [18] and CASP10 [19] competition and was among the two dominant tools in CASP11 [73].

The energy functions or scoring functions used in FBA are directed by micro-state interactions existing within known protein structures. These energy terms or functions are also termed as “Knowledge Based Potentials” [74]. FBA algorithms sought out to optimize these energy functions. Though and on one hand, FBA based algorithms have witnessed the most success in the biannual CASP competition by designing algorithms around the principle that certain local structures are favoured by local amino acid sequence. On the other hand, this limits their ability to search for alternate conformation of the proteins within a single run which reduces their probability of discovering a novel fold.

## 2.4 Secondary Structural Elements Based Approaches

Algorithms employing the use of SSEs for building protein models usually focus on assembling the core backbone of the protein with an exception of loop regions leading to model refinement protocols. BCL::FOLD [15] is one of such algorithms built with the objective to overcome the size and complexity limits faced by most approaches. In the later edition, restraints recovered from sparse NMR data were also incorporated in the pipeline aiding in rapid identification of protein topology [40]. This was benchmarked on protein data set upto a length of 565 containing both soluble and membrane proteins. The algorithm was tested on 20 CASP11 targets, out of which it was able to produce a GDT\_TS score of 30% on average for twelve [52]. The average GDT\_TS score was accounted for 36%. The study was conducted by using targets belonging to different categories offered by CASP, for example T0 (regular targets), TP (predicted residue-residue contacts) and TS (NMR-NOE restraints) etc. This study also pointed out that better structures were predicted for proteins dominated by  $\alpha$ -helix than  $\beta$ -strands. The prediction accuracy also decreased with the size of the protein.

Another algorithm based on the similar principle is SSThread [53]. It predicts contacting pairs of  $\alpha$ -helices and  $\beta$ -strands from experimental structures, secondary structure prediction and contact map predictions. The overlapping pairs are then assembled into a core structure leading to the prediction of loop regions. The contact pairing strategy employed by SSThread has been shown to be better in predicting  $\beta$ -strand pairs than all  $\alpha$  pairs.

## 2.5 Deep Learning Based Approaches

Quite recently neural network based deep learning approaches have seen a boom in the field of protein structure prediction.

So far, deep learning approaches for PSP have vastly been used as one of the component in the entire pipeline rather than implicitly being implemented as the driving force. Majority of its use revolve around prediction of residue-residue contacts, which are primarily derived through co-evolutionary approaches and/or by building sequence alignment profiles [75–77].

Recent work done by Al Quraishi [78] focused on building a pure deep learning based prediction approach. It was designed as a one step algorithm for prediction of protein structure relying on end-to-end differential deep learning strategy. The emphasis was laid on not using



any co-evolutionary data or information from existing templates for protein model construction. Instead, the algorithm relied on data derived solely from protein sequence in question and evolutionary profile of individual residues within the sequence. This method achieved state-of-the-art results as observed in case of *ab initio modelling* protocols.

Another tool that has shown prominence in CASP13 is DeepMind’s AlphaFold [79]. It uses a two step process for protein structure determination, which also involves the use of co-evolutionary profiles to guide model building. Through this methodology high-accuracy models were constructed for 24 out of 43 test proteins achieving a TM-score of 0.7 and above in the template-free modelling domain.

The community is still beginning to explore the benefits of deep learning approaches into PSP. The major step back that can be encountered by these techniques would be related to lack of availability of structural data. Since, these approaches are based on training the algorithm based on the certain patterns followed by available data. Structure prediction field has always been slower than the sequence equivalent which translates into lower availability of data that can be used to train the algorithm. Hence, though deep learning based approaches can be better implemented in sequence domain of protein biology, it will take other advances in structural biology to push forward deep learning based approaches. Other problem that such approaches can be prone to is over training of the algorithms.

### 3 Hybrid Approaches

With the advancements made in computational approaches to protein structure prediction, the line between individual methodology is diminishing. Now the structure prediction community is moving forward towards the use of “Hybrid Approaches”, which do not strictly rely on pure template based or template-free prediction criteria but on the amalgamation of both. Bhageerath [80] is one such homology/*ab initio* hybrid protocol. It is available in the form of a web-server called Bhageerath-H [37]. The main focus of the pipeline is on reduction of conformational search space. Out of thousands of predicted models, top 5 are selected based on physio-chemical metric (pcSM) scoring function (specific to this algorithm). Efficiency of this software was put to test by using CASP10 targets with promising prediction results. After the assessment of its shortcomings, an updated version was presented in CASP12 meeting as BhageerathH+ [81].

In another study, Quark [49] and fragment-guided molecular dynamic (FG-MD) were added to I-Tasser pipeline [11, 82] to improve on the existing protocol [33, 83]. The basic idea was to introduce *ab initio* generated structures from QUARK into LOMETS [84] to find any hit with existing homologous template with a good TM-score. Top hits are then passed into I-Tasser pipeline for atomic refinement to obtain a structure with low rmsd. This combination produced better results for FM targets in CASP10 and CASP11 experiments than QUARK alone [33, 85]. MULTICOM\_NOVEL approach is one more example of a hybrid algorithm which was constructed by combining various complementary structure prediction pipelines including MULTICOM server, I-Tasser, RaptorX [14], Rosetta etc.

Chunk-Tasser can also be put into this category as it utilizes both chunks of folded secondary structural fragments along with fold-recognition to assemble protein structures [50].

On similar grounds, an initiative was undertaken in 2014 to combine methods of the best known protein structure prediction techniques and to come up with a pipeline which could generate better structures. This initiative came to be known as WeFold, where 13 labs collaborated to merge their algorithms forming 5 major branches [86]. The outcome was promising and the authors of this study discussed on further improvements to be made in prediction protocols as a result of this ‘cooperation’ [86].

## 4 Evolution of CASP and its contribution

CASP has been a contributing factor for the work done in the field of computational protein structure prediction. It is a biennial competition being conducted for around two decades serving as a platform to judge the accuracy of prediction pipelines. It has grown overtime into a protein structure prediction platform to qualify prediction strategies coming under domains like template-based modelling, template-free modelling, refinement protocols, contact prediction etc. [12, 17, 87, 88].

To keep a track of advancement in PSP techniques, CASP prepares a list of unpredicted protein sequences in each category every two years. This provides an uniformity in assessing the advancement perceived in each area of structure prediction. The protein sequence list provided for blind testing of *ab initio* modelling approaches often constitutes of protein sequences with “soon to be released” structures. Best models are determined on the basis of several criteria, one of them being a local-global alignment score called GDT\_TS score (Global Distance Test) [89]. It calculates the  $C\alpha$  distance between residues from model and template protein at defined rmsd cut-off values. Henceforth determining both local and global similarities between two protein molecules.

The initial achievement in protein tertiary structure prediction was observed in CASP4, but mainly for small proteins ( $\leq 120$  residues). In later years, the *ab initio* prediction field remained stagnant for about a decade until the introduction of better contact prediction approaches in CASP11 competing pipelines with promising improvements in prediction accuracy [90]. Similar trend was observed in CASP12 with the inclusion of alignment-based contact prediction methods [91].

Recently conducted CASP13 demonstrated further improvement on average GDT\_TS score due to the employment of deep learning approaches in structure prediction [78]. This served as an encouragement to dig further into deep learning based approaches to solve the protein folding problem.

## 5 Conclusion

Template-based prediction in general are quicker than experimental methods, at least in providing initial spatial arrangement of the protein. One of the major drawback of these approaches is the redundancy of information, i.e., no new fold or family can be discovered as it relies on building models from existing structures. In addition, these methods fail to establish the structural integrity of a protein sequence with decreasing sequence or structure identity.

This review peeks into few methods and possibilities of free-modelling techniques developed and available for the prediction of protein structure. *Ab initio* protein structure prediction still bare influence from PDB structures for optimizing the parameters of protein folding. This information helps them reduce the conformational space sampling requirements by maximizing the efficiency of energy functions. Most of the algorithms are still directed by a combination of knowledge-based potentials and physics-based approaches [92].

To date free-modelling has been well adapted for protein sequences length of 150 residues or below [16] [92, 93]. Few instances have seen algorithms overdoing themselves and going beyond the length restrictions to predict structure for longer proteins. CASP11 witnessed major success in *ab initio* protein structure prediction for a structure of length 256 a.a. [90].

The major challenges faced by this field starts with finding an efficient way to explore the conformational search space as a protein sequence can fold into indefinite forms. Thereby, reducing the plausible folding possibilities to the best probable fold is a hard task to achieve.

The length limitation could indicate towards the design strategy of the algorithms, many of which rely on defining domain boundaries prior to structure prediction. But, this would not

be a very strong case of argument as most reliable predictions still lie under the length of 150 residue, though single domain boundary could expand upto a length 200 to 250 residues. The only exception to this case seen over the past years of CASP competition where a protein of length 256 a.a. was accurately predicted [90].

One could also argue that as a community, we are still at the domain-level of structure prediction given the length limitations. Algorithms have been built that specifically target solving the structure of single-domain proteins [57, 94] or choices of restraints that are limited to single-domain proteins only [56]. Many of the designed algorithms tend to be validated on a data set of single-domain proteins as well [75].

Even so, in the case of small protein structure predictions, not all of them deliver model close to its nativity. It is just that chances of having a good structure is more if the length of the protein is around 150 or less.

There is also the case of stagnancy in the field for the majority of years and been pointed out a lot. One of the example is the discussion provided by the end of CASP10 [95]. The study pointed out that the results from CASP10 lay closer to what were observed during CASP5. It also reflected on the fact that it might be due to the gradual increase in the complexity of CASP targets along with the inclusion of multi-domain targets provided for modelling.

Another drawback faced by few of these algorithms can be the run time which can vary a lot depending on the size of the protein and the internal functioning of the algorithm. With the inclusion of artificial intelligence, the time scale for modelling has been reduced to milliseconds but generally, a single prediction can take from somewhere from few minutes to hours to days for an algorithm to complete.

Most of the algorithms still rely on manual intervention to complete the runs and so the human error should also be considered.

The point of PSP is not just high accuracy structure determination but also to ascertain the basis behind this biological process (protein folding). Thereby, finally answering questions like “why good protein become faulty and cause disease”.

*De novo* protein structure prediction still requires a lot of improvement, but at the same time it promises a better prospect of structure prediction in future. It brings with it a hope of predicting newer folds at a faster pace when compared to experimental approaches which can remain stuck for years altogether due to numerous reasons. In general computational structure prediction techniques though have a room for improvement are still quick when compared to traditional approaches [16]. If considering Template-Based modelling approaches, few limitations still persist whereas *ab initio* approaches can move a step ahead and might help understanding the basic principles of protein folding [93].

## Acknowledgements

The authors are most thankful to Yves-Henri Sanejouand for critical reading of the manuscript. SD is thankful to Conseil Régional de La Réunion and Fonds Social Européen for providing a PhD scholarship under tier 234275, convention DIREDD/20161451. BO is thankful to Conseil Régional Pays de la Loire for support in the framework of GRIOTE grant.

## Conflict of interest

Authors declare no competing interests.

## Authors contributions

Conception of the work was done by SD, RS, FC and BO. SD, RS and BO performed literature data collection and analysis. All four authors SD, RS, FC and BO wrote the manuscript. All authors approved the final version of the article.

## References

- [1] James C Whisstock and Arthur M Lesk. Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, 36(3):307–340, 2003.
- [2] Gina Kolata. Trying to crack the second half of the genetic code. *Science*, 233:1037–1040, 1986.
- [3] Emmanuel Boutet, Damien Lieberherr, Michael Tognolli, Michel Schneider, and Amos Bairoch. UniProtKB/Swiss-Prot. *Methods in molecular biology (Clifton, N.J.)*, 406:89–112, 2007.
- [4] Peter W. Rose, Andreas Prlić, Ali Altunkaya, Chunxiao Bi, Anthony R. Bradley, Cole H. Christie, Luigi Di Costanzo, Jose M. Duarte, Shuchismita Dutta, Zukang Feng, Rachel Kramer Green, David S. Goodsell, Brian Hudson, Tara Kalro, Robert Lowe, Ezra Peisach, Christopher Randle, Alexander S. Rose, Chenghua Shao, Yi Ping Tao, Yana Valasatava, Maria Voigt, John D. Westbrook, Jesse Woo, Huangwang Yang, Jasmine Y. Young, Christine Zardecki, Helen M. Berman, and Stephen K. Burley. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, 45(D1):D271–D281, 2017.
- [5] Max F. Perutz, Michael G. Rossmann, Ann F. Cullis, Hilary Muirhead, Georg Will, and A. C. T. North. Structure of Haemoglobin: A Three-Dimensional Fourier Synthesis at 5.5-Å. Resolution, Obtained by X-Ray Analysis. *Nature*, 185(4711):416–422, 1960.
- [6] Xavier Morelli, Alain Dolla, Myrjam Czjzek, P. Nuno Palma, Francis Blasco, Ludwig Krippahl, Jose J.G. Moura, and Françoise Guerlesquin. Heteronuclear NMR and soft docking: An experimental approach for a structural model of the cytochrome c553-ferredoxin complex. *Biochemistry*, 39(10):2530–2537, 2000.
- [7] Dong Xu and Yang Zhang. Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. *Scientific reports*, 3:1895, 2013.
- [8] Ewen Callaway. The Revolution Will Not Be Crystallized. *Nature*, 525:172–174, 2015.
- [9] Christian B. Anfinsen. The Formation and Stabilization of Protein Structure. *Biochemical Journal*, 128(4):737–749, 1972.
- [10] Christian B Anfinsen. Principles that Govern the Folding of Protein Chains. *Science*, 181(4096):223–230, 1973.
- [11] Amrith Roy, Alper Kucukural, and Yang Zhang. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols*, 5(4):725–738, 2010.
- [12] Michael Feig. Computational protein structure refinement: almost there, yet still so far to go. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 7(3), 2017.

- [13] Andras Fiser. Template-based protein structure modeling. *Methods in molecular biology (Clifton, N.J.)*, 673:73–94, 2010.
- [14] Morten Källberg, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu, and Jinbo Xu. Template-based protein structure modeling using the RaptorX web server. *Nature Protocols*, 7(8):1511–1522, 2012.
- [15] Mert Karaka, Nils Woetzel, Rene Staritzbichler, Nathan Alexander, Brian E. Weiner, and Jens Meiler. BCL::Fold - De Novo Prediction of Complex and Large Protein Topologies by Assembly of Secondary Structure Elements. *PLoS ONE*, 7(11), 2012.
- [16] Ling-Hong Hung, Shing-Chung Ngan, and Ram Samudrala. *De Novo Protein Structure Prediction*, pages 43–63. Springer New York, New York, NY, 2007.
- [17] Moshe Ben-David, Orly Noivirt-Brik, Aviv Paz, Jaime Prilusky, Joel Sussman, and Yaakov Levy. Assessment of casp8 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, 77(Suppl. 9):50–65, 2009.
- [18] Lisa Kinch, Shuo Yong Shi, Qian Cong, Hua Cheng, Yuxing Liao, and Nick V. Grishin. CASP9 assessment of free modeling target predictions. *Proteins: Structure, Function, and Bioinformatics*, 79(Suppl. 10):59–73, 2011.
- [19] Chin Hsien Tai, Hongjun Bai, Todd J. Taylor, and Byungkook Lee. Assessment of template-free modeling in CASP10 and ROLL. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl. 2):57–83, 2014.
- [20] David Simoncini, Thomas Schiex, and Kam Y.J. Zhang. Balancing exploration and exploitation in population-based sampling improves fragment-based de novo protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 85(5):852–858, 2017.
- [21] Lisa Kinch, Wenlin Li, Richard Schaeffer, Roland Dunbrack, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Nick Grishin. Casp 11 target classification. *Proteins: Structure, Function and Bioinformatics*, 84:20–33, 2016.
- [22] Bee Yin Khor, Gee Jun Tye, Theam Soon Lim, and Yee Siew Choong. General overview on structure prediction of twilight-zone proteins. *Theoretical Biology and Medical Modelling*, 12(1):15, 2015.
- [23] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The SWISS-MODEL workspace: A web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, 2006.
- [24] Lorenza Bordoli, Florian Kiefer, Konstantin Arnold, Pascal Benkert, James Battey, and Torsten Schwede. Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols*, 4(1):1–13, 2008.
- [25] Alexander Miguel Monzon, Diego Javier Zea, Cristina Marino-Buslje, and Gustavo Parisi. Homology modeling in a dynamical world. *Protein Science*, 26:2195–2206, 2017.
- [26] Andras Fiser, Roberto Sánchez, Francisco Melo, and Andrej Fiser. Comparative Protein Structure Modeling. *Computational Biochemistry and Biophysics*, 2001.
- [27] Benjamin Webb and Andrej Sali. Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54(1):5.6.1–5.6.37, 2016.

- [28] Burkhard Rost, Reinhard Schneider, and Chris Sander. Protein fold recognition by prediction-based threading. *Journal of Molecular Biology*, 270(3):471–480, 1997.
- [29] Jeffrey Skolnick and Daisuke Kihara. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins: Structure, Function, and Bioinformatics*, 42(3):319–331, 2001.
- [30] William R Taylor and Inge Jonassen. A structural pattern-based method for protein fold recognition. *Proteins: Structure Function, and Bioinformatics*, 56(2):222–34, 2004.
- [31] Jinbo Xu, Feng Jiao, and Libo Yu. Protein structure prediction using threading. *Methods in molecular biology (Clifton, N.J.)*, 413:91–121, 2008.
- [32] David E. Kim, Ben Blum, Philip Bradley, and David Baker. Sampling Bottlenecks in De novo Protein Structure Prediction. *Journal of Molecular Biology*, 393(1):249–260, 2009.
- [33] Wenxuan Zhang, Jianyi Yang, Baoji He, Sara Walker, Hongjiu Zhang, Brandon Govindarajoo, Jouko Virtanen, Zhidong Xue, Hong-Bin Shen, and Yang Zhang. Integration of quark and i-tasser for ab initio protein structure prediction in casp11. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl. 1):76–86, 2015.
- [34] Debswapna Bhattacharya, Renzhi Cao, and Jianlin Cheng. UniCon3D: De novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. *Bioinformatics*, 32(18):2791–2799, 2016.
- [35] Paul IW de Bakker, Nicholas Furnham, Tom L. Blundell, and Mark A. DePristo. Conformer generation under restraints. *Current Opinion in Structural Biology*, 16(2):160–165, 2006.
- [36] Adam Liwo, Cezary Czaplewski, Stanisław Ołdziej and Harold A Scheraga. Computational techniques for efficient conformational sampling of proteins Adam. *Structure*, 18(2):134–139, 2008.
- [37] B Jayaram, Priyanka Dhingra, Avinash Mishra, Rahul Kaushik, Goutam Mukherjee, Ankita Singh, and Shashank Shekhar. Bhageerath-H: A homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. *BMC Bioinformatics*, 15(Suppl. 16):S7, 2014.
- [38] Peter M Bowers, Charlie E Strauss, and David Baker. De novo protein structure determination using sparse NMR data. *Journal of Biomolecular NMR*, 18(4):311–318, 2000.
- [39] Maya Topf, Matthew L. Baker, Marc A. Marti-Renom, Wah Chiu, and Andrej Sali. Refinement of protein structures by iterative comparative modeling and cryoEM density fitting. *Journal of Molecular Biology*, 357(5):1655–1668, 2006.
- [40] Brian Weiner, Nathan Alexander, Louesa Akin, Nils Woetzel, Mert Karakas, and Jens Meiler. Bcl::fold-protein topology determination from limited nmr restraints. *Proteins: Structure Function, and Bioinformatics*, 82:587–595, 2014.
- [41] James. U. Bowie and David Eisenberg. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proceedings of the National Academy of Sciences*, 91(10):4436–4440, 1994.
- [42] Kim T. Simons, Charles Kooperberg, Enoch Huang, and David Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–225, 1997.

- [43] Kim T. Simons, Rich Bonneau, Ingo Ruczinski, and David Baker. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins: Structure, Function, and Bioinformatics*, 37(Suppl. 3):171–176, 1999.
- [44] Richard Bonneau, Jerry Tsai, Ingo Ruczinski, Dylan Chivian, Carol Rohl, Charlie E M Strauss, and David Baker. Rosetta in CASP4: Progress in ab initio protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 45(Suppl. 5):119–126, 2001.
- [45] Srivatsan Raman, Robert Vernon, James Thompson, Michael Tyka, Ruslan Sadreyev, Jimin Pei, David Kim, Elizabeth Kellogg, Frank Dimaio, Oliver Lange, Lisa Kinch, Will Sheffler, Bong Hyun Kim, Rhiju Das, Nick V. Grishin, and David Baker. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics*, 77(Suppl. 9):89–99, 2009.
- [46] Hahnbeom Park, Frank DiMaio, and David Baker. Casp11 refinement experiments with rosetta. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl. 1):314–322, 2015.
- [47] Sergey Ovchinnikov, David Kim, Ray Wang, Yuan Liu, Frank DiMaio, and David Baker. Improved de novo structure prediction in casp11 by incorporating co-evolution information into rosetta. *Proteins: Structure, Function, and Bioinformatics*, 84:67–75, 2015.
- [48] Rebecca Faye Alford, Andrew Leaver-Fay, Jeliasko R. Jeliaskov, Matthew J O’Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V Shapovalov, P. Douglas Renfrew, Vikram Khipple Mulligan, Kalli Kappel, Jason W. Labonte, Michael Steven Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017.
- [49] Dong Xu and Yang Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics*, 80(7):1715–1735, 2012.
- [50] Hongyi Zhou and Jeffrey Skolnick. Ab initio protein structure prediction using chunk-TASSER. *Biophysical Journal*, 93(5):1510–1518, 2007.
- [51] S Ołdziej, C Czaplewski, A Liwo, M Chinchio, M Nancias, J A Vila, M Khalili, Y A Arnautova, A Jagielska, M Makowski, H D Schafroth, R Kaźmierkiewicz, D R Ripoll, J Pillardy, J A Saunders, Y K Kang, K D Gibson, and H A Scheraga. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7547–7552, 2005.
- [52] Axel W. Fischer, Sten Heinze, Daniel K. Putnam, Bian Li, James C. Pino, Yan Xia, Carlos F. Lopez, and Jens Meiler. CASP11 - An evaluation of a modular BCL: Fold-based protein structure prediction pipeline. *PLoS ONE*, 11(4), 2016.
- [53] Kevin J. Maurice. SSThread: Template-free protein structure prediction by threading pairs of contacting secondary structures followed by assembly of overlapping pairs. *Journal of Computational Chemistry*, 35(8):644–656, 2014.
- [54] Brinda Vallat, Carlos Madrid-Aliste, and Andras Fiser. Modularity of Protein Folds as a Tool for Template-Free Modeling of Structures. *PLoS Computational Biology*, 11(8):1–16, 2015.

- [55] Daisuke Kihara, Hui Lu, Andrzej Kolinski, and Jeffrey Skolnick. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proceedings of the National Academy of Sciences of the United States of America*, 98(18):10125–10130, 2001.
- [56] Yang Zhang, Andrzej Kolinski, and Jeffrey Skolnick. TOUCHSTONE II: a new approach to ab initio protein structure prediction. *Biophysical Journal*, 85(2):1145–64, 2003.
- [57] Mirco Michel, Sikander Hayat, Marcin J. Skwark, Chris Sander, Debora S. Marks, and Arne Elofsson. PconsFold: Improved contact predictions improve protein models. *Bioinformatics*, 30(17):482–488, 2014.
- [58] David Simoncini, Francois Berenger, Rojan Shrestha, and Kam Y J Zhang. A probabilistic fragment-based protein structure prediction algorithm. *PLoS ONE*, 7(7):1–11, 2012.
- [59] David Simoncini and Kam Y J Zhang. Efficient Sampling in Fragment-Based Protein Structure Prediction Using an Estimation of Distribution Algorithm. *PLoS ONE*, 8(7):1–10, 2013.
- [60] John L. Klepeis and Christodoulos A. Floudas. ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. *Biophysical journal*, 85(4):2119–2146, 2003.
- [61] Ashwin. Subramani, Yinan Wei, and Christodoulos A. Floudas. ASTRO-FOLD 2.0: An enhanced framework for protein structure prediction. *AIChE Journal*, 58(5):1619–1637, 2012.
- [62] Joe G. Greener, Shaun M. Kandathil, and David T. Jones. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. *Nature Communications*, 10(1):1–13, 2019.
- [63] Adam Liwo, Jooyoung Lee, Daniel R Ripoll, Jaroslaw Pillardy, and Harold A Scheraga. Protein structure prediction by global optimization of a potential energy function. *Proceedings of the National Academy of Sciences of the United States of America*, 96(10):5482–5485, 1999.
- [64] Corey Hardin, Taras V. Pogorelov, and Zaida Luthey-Schulten. Ab initio protein structure prediction. *Current Opinion in Structural Biology*, 12(2):176–181, 2002.
- [65] Alberto Perez, Joseph A Morrone, Emiliano Brini, Justin L Maccallum, and Ken A Dill. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances*, 2:e1601274, 2016.
- [66] Alpan Raval, Stefano Piana, Michael P. Eastwood, and David E. Shaw. Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations. *Protein Science*, 25(1):19–29, 2016.
- [67] Richard Bonneau and David Baker. Ab initio protein structure prediction: progress and prospects. *Annual Review of Biophysics and Biomolecular Structure*, 30(1):173–189, 2001.
- [68] Carlos Simmerling, Bentley Strockbine, and Adrian E. Roitberg. All-atom structure prediction and folding simulations of a stable protein. *Journal of the American Chemical Society*, 124(38):11258–11259, 2002.



- [69] Alpan Raval, Stefano Piana, Michael P. Eastwood, Ron O. Dror, and David E. Shaw. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. *Proteins: Structure, Function, and Bioinformatics*, 80(8):2071–2079, 2012.
- [70] Alpan Raval, Stefano Piana, Michael P. Eastwood, and David E. Shaw. Assessment of the utility of contact-based restraints in accelerating the prediction of protein structure using molecular dynamics simulations. *Protein Science*, 25(1):19–29, 2016.
- [71] Hai Nguyen, James Maier, He Huang, Victoria Perrone, and Carlos Simmerling. Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. *Journal of the American Chemical Society*, 136(40):13959–13962, 2014.
- [72] Fan Jiang and Yun Dong Wu. Folding of fourteen small proteins with a residue-specific force field and replica-exchange molecular dynamics. *Journal of the American Chemical Society*, 136(27):9536–9539, 2014.
- [73] Lisa Kinch, Wenlin Li, Bohdan Monastyrskyy, Andriy Kryshchak, and Nick Grishin. Evaluation of free modeling targets in casp11 and roll. *Proteins: Structure, Function, and Bioinformatics*, 84(Suppl. 1):51–66, 2015.
- [74] Evandro Ferrada and Francisco Melo. Effective knowledge-based potentials. *Protein Science*, 18(7):1469–1485, 2009.
- [75] Jörg Schaarschmidt, Bohdan Monastyrskyy, Andriy Kryshchak, and Alexandre Bonvin. Assessment of contact predictions in casp12: Co-evolution and deep learning coming of age. *Proteins: Structure, Function, and Bioinformatics*, 86(Suppl. 1):51–66, 2017.
- [76] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Computational Biology*, 13(1):1–34, 2017.
- [77] Shaun M. Kandathil, Joe G. Greener, and David T. Jones. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1092–1099, 2019.
- [78] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell Systems*, 8(4):292 – 301.e3, 2019.
- [79] Andrew Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577:1–5, 2020.
- [80] B. Jayaram, Priyanka Dhingra, Bharat Lakhani, and Shashank Shekhar. Bhageerath - Targeting the near impossible: Pushing the frontiers of atomic models for protein tertiary structure prediction. *Journal of Chemical Sciences*, 124(1):83–91, 2012.
- [81] Rahul Kaushik, Ankita Singh, Debarati Dasgupta, Amita Pathak, Shashank Shekhar, and B. Jayaram. BhageerathH+: A hybrid methodology based software suite for protein tertiary structure prediction. In *CASP12 Proceedings*, pages 25–26, 2016.
- [82] Jianyi Yang and Yang Zhang. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Research*, 43(W1):W174–W181, 2015.

- [83] Dong Xu, Jian Zhang, Ambrish Roy, and Yang Zhang. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins: Structure, Function, and Bioinformatics*, 79(Suppl. 10):147–160, 2011.
- [84] Sitao Wu and Yang Zhang. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35(10):3375–3382, 2007.
- [85] Yang Zhang. Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl. 2):175–187, 2014.
- [86] George A. Khoury, Adam Liwo, Firas Khatib, Hongyi Zhou, Gaurav Chopra, Jaume Bacardit, Leandro O. Bortot, Rodrigo A. Faccioli, Xin Deng, Yi He, Pawel Krupa, Jilong Li, Magdalena A. Mozolewska, Adam K. Sieradzian, James Smadbeck, Tomasz Wirecki, Seth Cooper, Jeff Flatten, Kefan Xu, David Baker, Jianlin Cheng, Alexandre C.B. Delbem, Christodoulos A. Floudas, Chen Keasar, Michael Levitt, Zoran Popović, Harold A. Scheraga, Jeffrey Skolnick, and Silvia N. Crivelli. WeFold: A competition for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 82(9):1850–1868, 2014.
- [87] Ralf Jauch, Hock Chuan Yeo, Prasanna R Kolatkar, and Neil D Clarke. Assessment of CASP7 structure predictions for template free targets. *Proteins: Structure, Function, and Bioinformatics*, 69(S8):57–67, 2007.
- [88] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP)—Round XII. *Proteins: Structure, Function, and Bioinformatics*, 86:7–15, 2018.
- [89] Adam Zemla. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13):3370–3374, 2003.
- [90] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI. *Proteins: Structure, Function, and Bioinformatics*, 84:4–14, 2016.
- [91] Luciano A. Abriata, Giorgio E. Tamò, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Matteo Dal Peraro. Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 86:97–112, 2018.
- [92] Yang Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, 2008.
- [93] Jooyoung Lee, Peter L Freddolino, and Yang Zhang. Ab initio protein structure prediction. In *From protein structure to function with bioinformatics*, pages 3–35. Springer, 2017.
- [94] David E. Kim, Dylan Chivian, and David Baker. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Research*, 32(WEB SERVER ISS.):526–531, 2004.
- [95] Andriy Kryshtafovych, Krzysztof Fidelis, and John Moult. CASP10 results compared to those of previous CASP experiments. *Proteins: Structure, Function, and Bioinformatics*, 82(Suppl. 2):164–174, 2014.