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

Surbhi Dhingra¹, Ramanathan Sowdhamini², Frédéric Cadet³,⁴, and Bernard Offmann*¹

¹Université de Nantes, CNRS, UFIP, UMR6286, F-44000 Nantes, France
²Computational Approaches to Protein Science (CAPS), National Centre for Biological Sciences (NCBS), Tata Institute for Fundamental Research (TIFR), Bangalore 560-065, India
³Sorbonne University Paris, UMR S1134, BIGR, Inserm, F-75015 Paris, France
⁴DSIMB, UMR S1134, BIGR, Inserm, Laboratory of Excellence GR-Ex, Faculty of Sciences and Technology, University of La Reunion, F-97715, Saint-Denis, 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 overtime, 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

*corresponding author : bernard.offmann@univ-nantes.fr
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 observed on 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

**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 atomic levels. The equivalent structural database, RCSB [4] (https://www.rcsb.org) documents around 160,000 structures 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 struc-
tures. 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 envi-
ronment." [10]. In other words, a protein attains its conformational nativity
when its environment is at its lowest Gibbs free energy levels. Another state-
ment 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, limit-
ing 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) [15,16]. A representa-
tive flowchart of the categorization is illustrated in Figure 1. This clas-
sification has been adopted by well-known biennial competition of protein
structure prediction, Critical Assessment of protein Structure Prediction
(CASP) [17–21]. Results from this competition benchmark the improvement
in the field of computational protein structure prediction [20,22]. Majority
of progress witnessed in this field is in construction of protein models us-
ing 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 similarities 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]. Struc-
tures 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 tem-
plate [23]. On the basis of shared sequence identity, it can be classified into Homology Modelling (HM) [24–26], Comparative Modelling (CM) [27, 28] and Threading approaches (fold-recognition) [29–32]. 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 [23]. 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 template 3D structure.

**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 homologs as template" [14,23,33–35]. 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 [36–38].

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 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 [23, 35]. 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. [39–41].
| Reference | Algorithms/Servers | Strategy/Approach |
|-----------|------------------|------------------|
| 42–49     | Rosetta          | Fragment-assembly using MC simulations, all-atom energy function to determine structure and clustering of models. |
| 7,50      | Quark            | Fragment-based protein assembly using REML simulation guided by Rosetta structure and clustering of models. |
| 60,19     | Edraose          | Fragment-based protein assembly using MC simulation all-atom energy function to determine structure and clustering of models. |
| 42,59     | DESTINI          | Evolution-based structure prediction pipeline using Ppense core. |
| 38,65     | Bhageerath-H     | Hybrid approach involving combination of several tools developed. |
| 35,52     | UniCon3D         | Using foldons and probabilistic models to capture local backbone structural preference and side chain conformation search space. |
| 38,65     | UNRES            | Physics-based conformational space search using UNRES energy function. |
| 38,65     | Touchstone       | Secondary and tertiary restraints prediction through threading. |
| 38,52     | BCL::FOLD        | Assembling secondary structural elements using MC sampling and knowledge-based energy functions. |
| 62,63     | Astro-Fold       | Folding amino acid sequences using first-principle based approaches. |
| 63,69     | Pfam-Fold        | Preferential-based all-atom energy function to produce unique models. |
| 60,19     | Edraose          | Fragment-based protein assembly using MC simulation all-atom energy function to determine structure and clustering of models. |
| 60,19     | DESTINI          | Evolution-based structure prediction pipeline using Ppense core. |
| 21        | Edraose          | Using foldons and probabilistic models to capture local backbone structural preference and side chain conformation search space. |
| 16,41,93  | BCL::FOLD        | Assembling secondary structural elements using MC sampling and knowledge-based energy functions. |
| 28        | Pfam-Fold        | Folding amino acid sequences using first-principle based approaches. |
| 38,65     | Bhageerath-H     | Hybrid approach involving combination of several tools developed. |
| 35,52     | UniCon3D         | Using foldons and probabilistic models to capture local backbone structural preference and side chain conformation search space. |
| 38,65     | UNRES            | Physics-based conformational space search using UNRES energy function. |
| 38,65     | Touchstone       | Secondary and tertiary restraints prediction through threading. |
| 38,65     | BCL::FOLD        | Assembling secondary structural elements using MC sampling and knowledge-based energy functions. |
| 62,63     | Astro-Fold       | Folding amino acid sequences using first-principle based approaches. |
| 63,69     | Pfam-Fold        | Preferential-based all-atom energy function to produce unique models. |
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 Cα-Cα 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 structure prediction approaches based on the core methodologies used to develop them.

**Physics Based**

These formed the basis of initial algorithms built under the emerging field. The main idea behind developing these physics-based approaches is to utilise no information from existing structures. The philosophy backing their design is to obtain lowest energy conformation model by folding the protein sequence using quantum mechanism and coulomb potential [65–67]. 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 [68,69]. 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 alongwith the meddling of erroneous force-field that restrict the model to attain it. [12,70–72].

MELD (Modelling Employing Limited Data) [68] 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 small proteins [68]. 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 [73]. 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) [74]. 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) [75].

**Fragment Based Approaches**

It is by far the most successful strategy used for template-free prediction of protein structure. This approach revolves around constructing a fragment library of varied lengths, each of which represents a pseudo-structure. The idea is to map information from protein fragments instead of using entire templates. Segments of query sequences are replaced by the fragment’s coordinates recorded in the fragment library. 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 [42]. 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 [42]. 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 [44, 45], 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.

SmotifsTF [56] 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 of average length 25 amino acids. QUARK [50] 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.

The energy functions or scoring functions used in FBA are directed by microstate interactions existing within known protein structures. These energy terms or functions are also termed as "Knowledge Based Potentials" [76].
Secondary Structural Elements Based Approach

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 [16] is one of such algorithms 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 [41].

Another algorithm based on the similar principle is SSThread [54]. 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.

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 [77] is one such homology/$ab\ initio$ hybrid protocol. It is available in the form of a web-server called Bhageerath-H [38]. The main focus of the pipeline is to reduce 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 released as BhageerathH+ [55].

In another study, Quark [50] and fragment-guided molecular dynamic (FG-MD) were added to I-Tasser pipeline [11, 78] to improve on the existing protocol [34, 79]. The basic idea was to introduce $ab\ initio$ generated structures from QUARK into LOMETS [80] 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 [34, 81]. MULTICOM_NOVEL approach is one more example of 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 [51].

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 [82]. 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 'coopetition' [82].

**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, template-free, refinement protocols, contact prediction etc. [12,18,83,84].

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 proteins with "soon to be released" structures. Best models are determined on the basis of a local-global alignment score called GDT_TS score (Global Distance Test) [85]. It calculates the $\mathrm{C}^\alpha$ 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). Later, the *ab initio* prediction field remained stagnant for a decade until the introduction of contact prediction in CASP11 competing pipelines with promising improvements in prediction accuracy. [86]. Similar trend was observed in CASP12 with the inclusion of alignment-based contact prediction methods [87].

Recently conducted CASP13 demonstrated further improvement on average GDT_TS score due to the employment of deep learning approaches in
structure prediction [88].

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 peaks 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 [89]. To date free-modelling has been been well adapted for protein sequences ranging upto 100-150 amino acids in length [17, 89, 90]. 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 [86].

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 [17]. 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 [90].

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,
Conflict of interest

Authors declare no competing interests.

References

[1] James C. Whisstock and Arthur M. Lesk. Prediction of protein function from protein sequence and structure, aug 2003.

[2] Gina Bari Kolata. Trying to crack the second half of the genetic code. Science, 233 4768:1037–9, 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] M. F. Perutz, M. 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 Czijzek, 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] Ambrish 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 Andras. 673:1–20, 2010.

[14] M Källberg, Haipeng Wang, Sheng Wang, Jian Peng, Zhiyong Wang, Hui Lu, and Jinbo Xu. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc.*, 7(8):1511–1522, 2012.

[15] Travis P Schrank. HHS Public Access. 22(7):95–121, 2016.

[16] 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.

[17] Ling-hong Hung, Shing-chung Ngan, and Ram Samudrala. De Novo Protein Structure Prediction. 2:43–63, 2007.

[18] Moshe Ben-david, Orly Noivirt-brik, Aviv Paz, Jaime Prilusky, Joel L Sussman, and Yaakov Levy. Assessment of CASP8 structure predictions for template free targets. (May):50–65, 2009.

[19] 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.
[20] 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.

[21] 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.

[22] Lisa N. Kinch, Wenlin Li, R. Dustin Schaeffer, Roland L. Dunbrack, Bohdan Monastyrskyy, Andriy Kryshtafovych, and Nick V. Grishin. CASP 11 Target Classification. *Proteins: Structure, Function, and Bioinformatics*, (January):n/a–n/a, 2016.

[23] 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.

[24] 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.

[25] 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.

[26] Alexander Miguel Monzon, Diego Javier Zea, Cristina Marino-Buslje, and Gustavo Parisi. Homology modeling in a dynamical world. *Protein Science*, 26:2195–2206, 2017.

[27] Andras Fiser, Roberto Sánchez, Francisco Melo, and Andrej Fiser. Comparative Protein Structure Modeling. *Computational Biochemistry and Biophysics*, 2001.

[28] N Eswar, B Webb, M A Marti-Renom, M S Madhusudhan, D Eramian, M Shen, U Pieper, and A Sali. Comparative protein structure modeling using Modeller, volume Chapter 5. 2006.

[29] Burkhard Rost, Reinhard Schneider, and Chris Sander. Protein fold recognition by prediction-based threading. *Journal of Molecular Biology*, 270(3):471–480, 1997.
[30] J Skolnick and D Kihara. Defrosting the frozen approximation: PROSPECTOR—a new approach to threading. *Proteins*, 42(3):319–331, feb 2001.

[31] William R Taylor and Inge Jonassen. A structural pattern-based method for protein fold recognition. *Proteins*, 56(2):222–34, 2004.

[32] Jinbo Xu, Feng Jiao, and Libo Yu. Protein structure prediction using threading. *Methods in molecular biology (Clifton, N.J.)*, 413:91–121, 2008.

[33] 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.

[34] Wenxuan Zhang, Jianyi Yang, Baoji He, Sara Elizabeth 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*, (August):76–86, 2015.

[35] 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.

[36] 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.

[37] Adam Liwo, Cezary Czaplewski, Stanislaw Oldziej and Harold A Scheraga. Computational techniques for efficient conformational sampling of proteins Adam. *Structure*, 18(2):134–139, 2008.

[38] B Jayaram, Priyanka Dhindra, 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.

[39] P M Bowers, C E Strauss, and D Baker. De novo protein structure determination using sparse NMR data. *Journal of biomolecular NMR*, 18(4):311–318, 2000.
[40] 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.

[41] Brian E. Weiner, Nathan Alexander, Louesa R. Akin, Nils Woetzel, Mert Karakas, and Jens Meiler. BCL::Fold – Protein topology determination from limited NMR restraints. 82(4):587–595, 2015.

[42] J. U. Bowie and D. 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.

[43] 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.

[44] 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 Genetics*, 37(SUPPL. 3):171–176, 1999.

[45] 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 Genetics*, 45(SUPPL. 5):119–126, 2001.

[46] 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.

[47] Hahnbeom Park, Frank Dimaio, and David Baker. CASP11 refinement experiments with ROSETTA. *Proteins: Structure, Function and Bioinformatics*, (May):314–322, 2015.

[48] Sergey Ovchinnikov, David E. Kim, Ray Yu Ruei Wang, Yuan Liu, Frank Dimaio, and David Baker. Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta.
Proteins: Structure, Function and Bioinformatics, (November):67–75, 2016.

[49] Rebecca Faye Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, 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, page acs.jctc.7b00125, 2017.

[50] 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.

[51] H Zhou and J Skolnick. Ab initio protein structure prediction using chunk-TASSER. Biophys J, 93(5):1510–1518, 2007.

[52] S Oldziej, C Czaplewski, A Liwo, M Chinchio, M Nnias, J A Vila, M Khalili, Y A Arnaoutova, 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.

[53] 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.

[54] 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.

[55] Shashank Shekhar Rahul Kaushik, Ankita Singh, Debarati DasGupta, Amita Pathak and B. Jayaram. BhageerathH+: A hybrid methodology based software suite for protein tertiary structure prediction. CASP12... (April 2017):4–6, 2016.
[56] 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.

[57] D Kihara, H Lu, a Kolinski, and J 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.

[58] 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.

[59] Mirco Michel, Sikander Hayat, Marcin J. Skwark, Chris Sander, Deobra S. Marks, and Arne Elofsson. PconsFold: Improved contact predictions improve protein models. *Bioinformatics*, 30(17):482–488, 2014.

[60] 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.

[61] 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.

[62] JL Klepeis and CA 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.

[63] A. Subramani, Y. Wei, and C. A. Floudas. ASTRO-FOLD 2.0: An enhanced framework for protein structure prediction. *AIChE Journal*, 58(5):1619–1637, 2012.

[64] Mu Gao, Hongyi Zhou, and Jeffrey Skolnick. DESTINI: A deep-learning approach to contact-driven protein structure prediction. *Scientific Reports*, 9(1), dec 2019.

[65] A Liwo, J Lee, D R Ripoll, J Pillardy, and H A Scheraga. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci.*, 96(10):5482–5485, 1999.

[66] Corey Hardin, Taras V. Pogorelov, and Zaida Luthey-Schulten. Ab initio protein structure prediction. *Current Opinion in Structural Biology*, 12(2):176–181, 2002.
[67] Peter L. Freddolino Jooyoung Lee and Yang Zhang. *Ab Initio Protein Structure Prediction*. 2017.

[68] 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.

[69] 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.

[70] Richard Bonneau and David Baker. Ab initio protein structure prediction: progress and prospects. *Annual review of biophysics and biomolecular structure*, 30(August 2017):173–89, 2001.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] Evandro Ferrada and Francisco Melo. Effective knowledge-based potentials. *Protein Science*, 18(7):1469–1485, 2009.

[77] B. Jayaram, Priyanka Dhangra, 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.

[78] Jianyi Yang and Yang Zhang. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Research*, 43(W1):W174–W181, 2015.

[79] 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.

[80] Sitao Wu and Yang Zhang. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35(10):3375–3382, 2007.

[81] 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.

[82] 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. Sieradzan, James Smadbeck, Tomasz Wirecki, Seth Cooper, Jeff Flatten, Kefan Xu, David Baker, Jianlin Cheng, Alexandre C.B. Delbem, Christodoulous A. Floudas, Chen Keasar, Michael Levitt, Zoran Popović, Harold A. Scheraga, Jeffrey Skolnick, and Silvia N. Crivelli. WeFold: A coopepetition for protein structure prediction. *Proteins: Structure, Function and Bioinformatics*, 82(9):1850–1868, 2014.

[83] 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.

[84] 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, mar 2018.

[85] Adam Zemla. LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Research*, 31(13):3370–3374, jul 2003.
[86] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Anna Tramontano. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins: Structure, Function and Bioinformatics*, (April):4–14, 2016.

[87] Luciano A. Abriata, Giorgio E. Tamò, Bohdan Monastyryskyy, 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, mar 2018.

[88] Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell systems*, 8(4):292–301, 2019.

[89] Yang Zhang. Progress and challenges in protein structure prediction. *Current Opinion in Structural Biology*, 18(3):342–348, 2008.

[90] Jooyoung Lee, Sitao Wu, and Yang Zhang. Ab initio protein structure prediction. *From Protein Structure to Function with Bioinformatics*, pages 3–25, 2009.