Artificial intelligence for template-free protein structure prediction: a comprehensive review

Artificial Intelligence for Template-Free Protein Structure Prediction: A Comprehensive Review

M. M. Mohamed Mufassirin[1,3*†], M. A. Hakim Newton[2,4†] and Abdul Sattar[1,2]

[1*]School of ICT, Griffith University, 170 Kessels Road, Nathan, Brisbane, 4111, Queensland, Australia.
[2]IIIS, Griffith University, 170 Kessels Road, Nathan, Brisbane, 4111, Queensland, Australia.
[3]Department of Computer Science, South Eastern University of Sri Lanka, Main Street, Sammanthurai, Ampara, 32200, Eastern Province, Sri Lanka.
[4]School of Information and Physical Sciences, The University of Newcastle, University Drive, Callaghan, 2308, New South Wales, Australia.

*Corresponding author(s). E-mail(s): m.mufassirin@griffithuni.edu.au;
Contributing authors: mahakim.newton@newcastle.edu.au; a.sattar@griffith.edu.au;
†These authors contributed equally and are joint first authors.

Abstract

Protein structure prediction (PSP) is a grand challenge in bioinformatics, drug discovery, and related fields. PSP is computationally challenging because of an astronomically large conformational space to be searched and an unknown very complex energy function to be minimised. To obtain a given protein's structure, template-based PSP approaches adopt a similar protein's known structure, while template-free PSP approaches work when no similar protein's structure is known. Currently, proteins with known structures are greatly outnumbered by proteins with unknown structures. Template-free PSP has obtained significant progress

recently via machine learning and search-based optimisation approaches. However, very accurate structures for complex proteins are yet to be achieved at a level suitable for effective drug design. Moreover, *ab initio* prediction of a protein's structure only from its amino acid sequence remains unsolved. Furthermore, the number of protein sequences with unknown structures is growing rapidly. Hence, to make further progress in PSP, more sophisticated and advanced artificial intelligence (AI) approaches are needed. However, getting involved in PSP research is difficult for AI researchers because of the lack of a comprehensive understanding of the whole problem, along with the background and the literature of all related sub-problems. Unfortunately, existing PSP review papers cover PSP research at a very high level and only some parts of PSP and only from a particular singular viewpoint. Using a systematic approach, this review paper provides a comprehensive survey of the state-of-the-art template-free PSP research to fill this knowledge gap. Moreover, covering required PSP preliminaries and computational formulations, this paper presents PSP research from AI perspectives, discusses the challenges, provides our commentaries, and outlines future research directions.

Keywords: Bioinformatics, Protein Structure Prediction, Machine Learning, Deep Learning, Search-based Optimisation

1 Introduction

Proteins are the building blocks of life. Proteins fold into three-dimensional (3D) structures during synthesis. A protein's function depends on its *native structure* that has the minimum free energy. Misfolded proteins might cause acute diseases. Also, a virus protein's function could be harmful to our body. A disease protein's function could be inhibited if a drug molecule could dock on the protein. Accurate knowledge of protein structures is, therefore, crucial in disease prediction, drug design, and related fields [1]. However, not much is known about the native structures of the proteins. So far, in-vitro techniques such as Nuclear Magnetic Resonance (NMR) [2] and X-ray crystallography [3] have been used in protein structure determination. Unfortunately, these methods are very time-consuming and failure-prone [4, 5]. To date, the structures of about 0.19 million proteins have been deposited in the Protein Data Bank (PDB) [6], accounting for only less than 0.1% of the total sequences in the UniProt database [7]. Consequently, the gap between the number of proteins already sequenced and the number of proteins of which native structures are known is hugely increasing. To aid the protein structure determination process, with the help of high-performance computers, computational protein structure prediction (PSP) approaches have been developed over the years [8, 9]. For the given proteins, PSP approaches strive to find selected decoy structures that could be further investigated to obtain the native structures. PSP is a grand challenge in bioinformatics, biotechnology, computational biology,

and computational chemistry. Computer Scientist Donald Knuth said in 1993, "Biology easily has 500 years of exciting problems to work on." With great computational challenges involved, PSP is certainly one such exciting problem.

Computational PSP obtained its propulsion when Anfinsen's classic study in the 1970s showed that the native structure of a protein could be determined solely by its amino acid sequence [10]. Since then, understanding the sequence-based paradigm has become a key component of contemporary biomedical research. Consequently, nucleotide sequences in the GenBank database [11] started getting translated into amino acid sequences and deposited in UniProt database [7]. Despite the extensive sequence data collection effort, not much progress has been observed in understanding each protein's biological functions. This is because protein functions are largely determined by their 3D structures, and in this regard, amino acid sequences directly alone provide only limited information. Interestingly, sequence information has been later used by machine-learning methods to predict structurally interacting residues [12–15]. This has then indirectly paved the way for dramatic improvement in structure prediction. Moreover, enhanced protein energy functions [16, 17] have allowed researchers to start with an approximate structure prediction model and refine it using an energy-guided refinement process to get it closer to the experimentally determined structure [18, 19]. Recently deep learning-based methods [20–24] have shown great progress and success in 3D PSP. Also, over the years, many online PSP servers have been available [20, 25–34].

Computational methods for PSP are divided into two major categories: Template-Based Modelling (TBM) and Template-Free Modelling (TFM) [35, 36]. TBM approaches predict structures of given protein sequences using known structures of selected template proteins. Template proteins are selected by *homology* or *comparative approaches* using similarity between amino acid sequences. Homology-based approaches work well when the given protein has at least 30% sequence similarity with the template protein [36, 37]. Template proteins can also be selected by *threading* or *fold recognition approaches* using structural similarity. We consider TBM approaches to be out of scope.

When template proteins are not found, we have to construct protein structures from scratch exploiting amino acid sequences more explicitly. This procedure has many different names such as *ab initio* modelling [38, 39], *de novo* modelling [40] and template-free modelling (TFM) [41]. In this paper, we use the term TFM uniformly to avoid confusion. TFM approaches are based on Anfinson's hypothesis that the native structure of a protein is determined only by the protein's amino acid sequence [10]. However, the energy function that drives the folding of a protein's amino acid sequence to its 3D native structure is not known precisely so far. Consequently, PSP approaches that do not utilise any knowledge from other proteins but solely rely only on the given protein's amino acid sequence and use all-atomic levels energy functions such as CHARMM [42] and AMBER [43] do not scale well with large and complex amino acid sequences. So, in recent times, machine learning models have been first trained to capture generalised knowledge from the amino acid sequences

and 3D structures of known proteins. In order to avoid over-fitting of the machine learning models, these known proteins should have very low sequence similarity (typically less than 25-30%) with given unknown proteins but still should allow capturing of the underlying mutual interaction patterns among the residues in the training proteins. Nevertheless, after training, the machine learning models are then used to directly predict the protein structures from their sequences or refine the already predicted protein structures.

TFM approaches have shown great progress recently by methods such as AlphaFold [21, 24], RoseTTAFold [23], trRosetta [20], and RaptorX-3DModeling [22]. Note that some of these methods [23, 24] also have integrated TBM approaches at some stages of their PSP pipelines. Nevertheless, AlphaFold2 [24] has demonstrated prediction accuracy competitive with experimental structures in a large number of cases and also has opened access to a much larger number of predicted protein structures [44]. After all these, very accurate PSP has now become available, and good accuracy could also be achieved for challenging proteins without having templates or similar proteins. Certainly, this recent progress would translate into enhancing homology search by treating highly accurately predicted structures like native structures. Also, the recent progress will lead to large-scale assignments of functions to proteins in a putative fashion. However, drug designing experts are particularly concerned that the predicted protein structures at the current accuracy levels are not yet reliable in the active and allosteric pocket areas where drug molecules could dock [45]; so further targeted improvement is clearly needed. Moreover, once very highly accurate PSP is possible with the help of the knowledge extracted from other proteins in one way or another, the very basic hypothesis that "the amino acid sequence of a protein alone determines its 3D native structure" should come back to have a greater focus on quenching the thirst of a theoretical and more sound scientific understanding and perhaps also to obtain even higher accuracy levels. Besides the biological concerns, from the computational point of view, the success of the current PSP methods is driven largely by the enormous computational resources that are beyond the reach of most PSP researchers. Furthermore, from artificial intelligence (AI) perspective, current PSP methods depend heavily on the voluminous training data (e.g. AlphaFold2's training proteins). At the same time, they should rely more on informed exploitation of knowledge for greater scaling.

From our long expertise both in core AI techniques and application of AI in PSP and other protein-related bioinformatics problems, we envisage further major progress in PSP would come from how well we adopt sophisticated and advanced AI methodologies in PSP. However, data-driven machine learning-based AI methods are often taken as black boxes, understanding that the more the data, the better the performance as the black box will magically and implicitly capture the underlying input-output relationships. The easy access to complex machine learning programs and also to supporting high-performing hardware has somewhat led to this situation. Unfortunately, complex models need more data, and more data needs more complex models—these two

thus have some kind of spiralling effect on each other. However, machine learning methods strive to achieve generality over the training examples and consequently, in general, lose accuracy. Moreover, covering all possible cases within a given finite training dataset is impossible. Therefore, as more data are available, repeated training of a machine learning algorithm with a massive dataset that includes both old and new data would be the way ahead. While AlphaFold2 [24] has shown great success in achieving high-quality predictions, its massive utilisation of almost all available protein data (both known and unknown proteins) along with its vast computational resources, raises computational and AI concerns. In the desperate need of a PSP solution, throwing everything is a reasonable choice, but once a working solution is somewhat obtained, perhaps it is time to start looking at the PSP area from a more critical point of view. To put things into perspective, perhaps some greater focus should be put on obtaining more efficient and simpler PSP methods that, of course, do not lose accuracy levels. Just to give an example, a Nature article [46] shows that a simple two-parameter logistic regression (that is, one neuron) obtains the same performance as that of a 13,451-parameter DNN in the seismic aftershock prediction area. In this context, we note that knowledge-driven AI methods could also play a vital role and offer alternative research avenues for PSP. Knowledge-driven AI methods strive to obtain performance using as little data as possible and rely on a clearer, crisper, and more explicit understanding of the underlying system. For example, abstraction-based AI methods could help tackle the scale of the problem, explanation-based AI methods could help reduce the dependency on huge data, reinforcement learning approaches could help deal with unknown energy functions, and constraint-guided approaches could help perform better sampling of the conformational search space.

Based on the above discussion, it is perhaps clear that PSP research would greatly benefit if AI researchers could inject their expertise more directly into the area. However, getting involved in PSP research is difficult for AI researchers because of the lack of opportunity to obtain a quick but comprehensive understanding of the whole problem, along with the background and the literature of all related sub-problems. Review and survey papers could help in this regard. Recent review and survey papers cover neuralisation of PSP [47, 48] and deep learning approaches for protein structure annotations [36, 49, 50]. Unfortunately, these review and survey papers cover only some parts of PSP and thus provide only a partial picture of PSP, again only from a particular singular viewpoint of deep learning and still at a very high level. This review paper provides a comprehensive survey of the template-free PSP literature. To obtain a systematic survey, we have used two popular bibliometric analysis tools Gephi [51] and Bibexcel [52]. Also, we have performed literature searching based on different combinations of keywords from our long experience in the PSP area. Covering required PSP preliminaries and computational formulations, in this review paper, we present PSP research from AI perspectives and then discuss the challenges, provide our commentaries, and outline future research directions.

The rest of the paper is organised as follows: we compare this review paper with other existing review papers; we then describe our survey methodology; next, we briefly cover preliminary knowledge of proteins and machine learning and search-based optimisation approaches; then, we provide an overview of PSP sub-problems; next, for each sub-problem, we cover its literature and discuss the challenges ahead; next, we provide an overall discussion; and finally we present our conclusions.

2 Comparison with Existing Review Papers

Template-free computational PSP research has been explored at time points. To show the relevance of this review paper, we give an overview of very recent review and survey papers, particularly those from the last three years.

- A review of deep learning methods for secondary structure prediction [49].
- An overview of fragment assembly, molecular dynamics, residue-residue contact prediction, and deep learning for PSP [53].
- A review [36] of physics and fragment assembly based approaches along with the end-to-end approach [54] and AlphaFold [21] .
- A survey [50] of prediction methods for secondary structure, solvent accessibility, contact maps, and backbone angles.
- A review [55] of RaptorX-contact [13], trRosetta [20], and AlphaFold2 [24] and deep learning for contact and distance maps.
- A historical overview [56] of molecular dynamics, fragment assembly, gradient descent method, and residue-residue constraints.
- A survey [48] of deep learning methods for fragment assembly and contact map-based PSP approaches.
- A review [47] of neuralisation of the PSP pipeline. This review has shown that in recent times computations based on energy models and sampling procedures have been gradually replaced by complex DNNs.
- A review [57] of computational methods used in PSP from mass spectrometry data such as hydroxyl radical protein footprinting, limited proteolysis, chemical cross-linking, hydrogen-deuterium exchange, ion mobility, and surface induced dissociation.
- A review [58] of recent deep learning methods applied in 3D structural proteomics and interactome modelling.

As we see from the above list of recent review and survey papers, most of these present PSP from a bioinformatics point of view and cover only subareas of PSP. A large majority of these papers are on deep learning methods since such PSP methods have produced excellent results in recent times. In this review paper, we view template-free PSP from AI perspectives and comprehensively cover the related subareas. We duly acknowledge the recently obtained great progress in template-free PSP and then perform a critical review and discuss potential research directions.

3 Systematic Literature Survey Methodology

In this review paper, we have utilised Gephi [51] and Bibexcel [52] software to conduct a comprehensive survey of the template-free PSP literature. We have used Scopus and the Web of Science databases with a rich set of 37 keywords in various combinations to collect and collate relevant papers published within the field of PSP. While searching for papers, we have used bioinformatics, structural biology, computational chemistry, and other similar words as subjects. After the search, we found 1507 papers that were published between 2014 and April 2021. Next, we reviewed the titles and keywords of the 1507 papers to determine whether the respective papers are primarily concerned with PSP or not. This first-round filtering process resulted in 792 papers on PSP. Next, another round of filtering was performed by reading the abstracts of the 792 papers. After the second round of filtering, we obtained 423 papers. Finally, the entire contents of the 423 papers have been comprehensively reviewed to check whether the topic of the paper is truly related to PSP. After this final filtering round, only 225 papers were deemed truly related to PSP. The filtering rounds have been performed using Bibexcel. Next, we used Gephi to perform co-citation analysis and obtain a bibliometric network along with different groups of papers. Fig. 1 shows a sample bibliographic network in which large dots represent high-impact papers (in terms of the number of papers that cite a paper), and colours represent groups of papers. Further to the systematic approach described so far, using our long experience in PSP, we have manually traced cited-by and cited-in papers for important papers known to us as notable recent contributions in template-free PSP research.

As described above, we have followed a rigorous survey methodology to explore the template-free PSP literature. This process represents a highly novel aspect of the analysis since existing literature reviews in PSP, to the best of our knowledge, have not described any systematic approach. It is worth noting that the application of bibliometric network analysis has been rare for reviewing PSP research. While existing review and survey papers explore about 100 papers, in this review paper, we explore over 200 papers. Of course, we cannot assert the inclusion of every paper in template-free PSP, but we have made a reasonable and human-assisted semi-automated methodological attempt to include high-impact and recent influential papers.

4 Preliminary Knowledge of PSP and AI

We provide preliminary concepts of protein structure prediction. We also provide brief overviews of machine learning approaches and search-based optimisation algorithms. Expert readers can skip some parts of this section.

4.1 Protein Structure Prediction

Proteins are sequences of amino acids. Regular proteins typically have 20 types of amino acids. The amino acid types differ in various characteristics, such

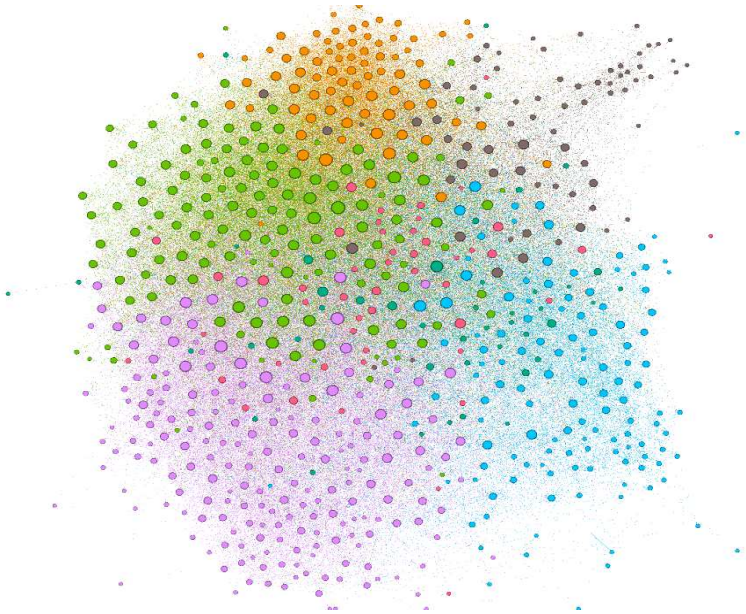

Fig. 1 Co-citation network diagram generated by Gephi. The size of a dot indicates the impact of the associated paper in terms of its citation, and the colour indicates the similarity of papers in terms of content.

as structure, size, shape, charge, hydrogen bonding capacity, hydrophobicity, and reactivity. Nevertheless, not each protein might have all of the 20 types of amino acid. Moreover, any amino acid can appear in a protein any number of times in any order subject to the stoichiometric constraints [59]. Constraints in PSP are also called *restraints*.

Fig. 2 shows only 4 of the 20 types of amino acids. As can be seen from the figure, each *amino acid* has a central carbon atom named *alpha carbon* (C^{α}). The C^{α} atom is connected to an *amino group* (NH_2) and a *carboxyl group* ($COOH$). A *side chain* could also be connected to the C^{α} atom except in one amino acid named Glycine. Essentially, each type of amino acid has a distinct side chain. The side chains, could have further carbon atoms such as *beta carbons* (C^{β}), *gamma carbons* C^{γ}), etc. and also other atoms such as sulphur (S), nitrogen (N), oxygen (O), and hydrogen (H). Some hydrogen atoms are not shown in Fig. 2. For convenience, sometimes, a side chain in an amino acid chain is generally denoted by R and is shown to be connected to C^{α}.

Amino acid types are denoted by unique single and three-letter codes. So a protein can essentially be represented by a sequence of those codes. Table 1 shows single- and three-letter amino acid codes and a protein's amino acid sequence using single-letter amino acid codes.

Given the sequence of amino acids in a protein, any two successive amino acids in the sequence are connected to form a *peptide bond*. Fig. 3 shows the formation of such a peptide bond when two amino acids are joined with each other. Considering the peptide bonds between all pairs of successive amino

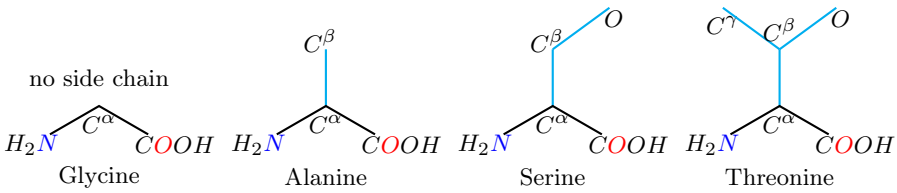

Fig. 2 4 amino acids with central C^α atoms, amino (NH_2) and carboxyl ($COOH$) groups, and side chains

Table 1 20 types of amino acid with their 3- and 1-letter codes, and the amino acid sequence of protein $1CRN$ using the 1-letter codes of the amino acids

Amino Acid	Codes		Amino Acid	Codes		Amino Acid	Codes		Amino Acid	Codes	
Alanine	ALA	A	Glutamine	GLN	Q	Leucine	LEU	L	Serine	SER	S
Arginine	ARG	R	Glutamic Acid	GLU	E	Lysine	LYS	K	Threonine	THR	T
Asparagine	ASN	N	Glycine	GLY	G	Methionine	MET	M	Tryptophan	TRP	W
Aspartic Acid	ASP	D	Histidine	HIS	H	Phenylalanine	PHE	F	Tyrosine	TYR	Y
Cysteine	CYS	C	Isoleucine	ILE	I	Proline	PRO	P	Valine	VAL	V

$1CRN =$ "$TTCCPSIVARSNFNVCRLPGTPEAICATYTGCIIIPGATCPGDYAN$"

acids in a protein, we obtain a *polypeptide chain*, which is also called the *main chain* or the *backbone* of the protein. The start and end terminals of a polypeptide chain are different [60]. One is the amino group, and the other is the carboxyl group, and so are respectively called the *N-terminal* and the *C-terminal*. These two terminals are shown in the right molecule in Fig. 3. Nevertheless, after the formation of a peptide bond, whatever is left from an amino acid in the backbone of the protein is called an amino acid residue or just *residue*. The length of a protein is the number of residues in the protein and is denoted by L.

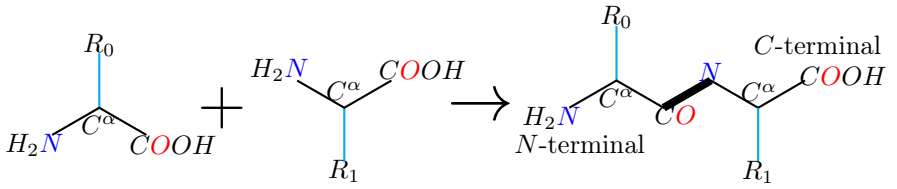

Fig. 3 Formation of a peptide bond (the thick line between the C and N atoms at the rightmost molecule) between each two successive amino acids of a protein, and the N- and C-terminals of the chain formed by the bond

For each amino acid, certain bonds in its molecular structure are rotatable in the 3D space. For example, the bond between N and C^α atoms and that between C^α and C atoms are rotatable. Side chains could also have rotatable bonds, but for brevity, we do not list them. For any rotatable bond, the bond acts as the axis of rotation and the atom at one end of the bond acts as the origin of rotation, and then any other atoms in the same amino acid residue or all atoms in other amino acid residues in the protein are rotated. In this way,

we obtain a 3D structure or *conformation* of a protein. Note that the peptide bonds between successive amino acids are not rotatable. Peptide bonds rather form planes comprising C^{α} and C atoms of the preceding amino acid residues and N and C^{α} atoms of the succeeding amino acid residues. Fig. 4 shows the 3D structure of a protein in various ways.

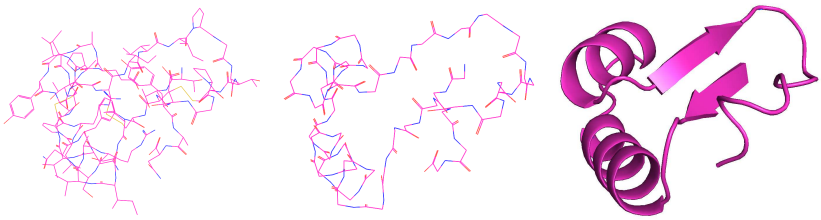

Fig. 4 Various representations of the 3D structure of protein $1CRN$: main and side chains (left), only main chain (middle), and secondary structures such as helices, sheets, and coils as cartoons (right)

Notice that in Fig. 4, some parts of the 3D structures exhibit local structures such as helices (or α helices), sheets (or β sheets), and coils (or loops). These local structures are called the *secondary structures* (SS) and considering variants of the three main types, and there are actually eight types of secondary structure in Dictionary of Secondary Structure of Proteins (DSSP) classification [61]. Fig. 5 shows that several secondary structures can build various super secondary structures such as parallel sheets or helices, and several super secondary structures can build various secondary structure motifs such as cylindrical shapes or bundles. These constructs essentially provide various abstraction levels. In this context, note that the entire 3D structure of a protein is called its *tertiary structure* while just the chain of amino acids is called the *primary structure*. Also, note that several tertiary structures of multiple polypeptide chains in a protein can form further bonds together and build a more complex structure called *quaternary structure*. Quaternary structures are out of the scope of this review paper.

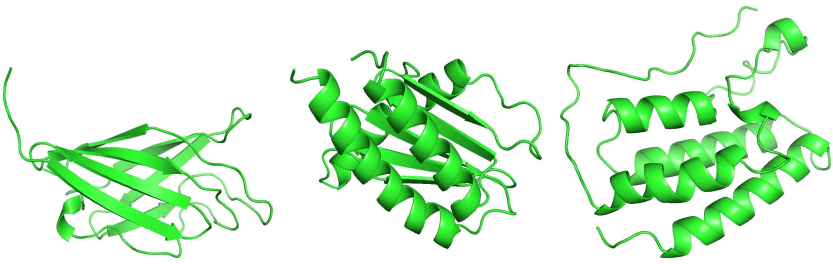

Fig. 5 Various super secondary structures and secondary structure motifs: parallel sheets and a cylindrical motif (left), parallel helices and parallel sheets (middle), and a bundle of parallel helices (right)

4.2 Search-Based Optimisation

Search-based optimisation involves finding *solutions* represented by *values* for given *variables* such that given *constraints* are satisfied and given *objective* functions are optimised. As shown in Fig. 6, search procedures could be constructive or perturbative. *Constructive search* starts from *empty solutions* having no variables assigned and within a loop extends current *partial solutions* by assigning to more variables. In each lap of the loop, extended new partial solutions are evaluated, and the best ones are accepted as the current partial solutions for the next lap. Constructive search algorithms stop when *complete solutions* having all variables assigned are obtained. In the PSP context, the constructive search could mean constructing conformations by iteratively adding one or more residues at a time and heuristically optimising the partial conformation. I-TASSER [26] algorithm, somewhat using a constructive approach, assembles fragments to obtain a full-length model. The use of complete constructive search techniques in PSP is extremely rare. We do not further explore the constructive search for PSP, but such approaches could be interesting as proteins naturally start getting folded as soon as their sequences start getting synthesised. *Perturbative search* starts from complete solutions along with their evaluations and, within a loop, perturbs the current solutions by changing the values of some variables. The new solutions obtained after perturbation are called *neighbour solutions*. In each lap of the loop, the neighbour solutions are evaluated, and the best ones are accepted as the current solutions for the next lap. Perturbative search algorithms run until a given termination criterion is satisfied. Most existing search-based PSP methods are perturbative by nature.

Both constructive and perturbative search algorithms could be of exponential complexities if they adopt *systematic* or *brute-force* approaches and enumerate all possible solutions. Upon practical considerations, heuristic and metaheuristic search approaches are therefore used. Heuristic and metaheuristic algorithms typically sample a subset of all possible solutions and thus do not guarantee to find any globally optimal solutions, if any. However, they are normally swift in finding high-quality solutions [62]. Nevertheless, *heuristic search approaches* are often greedy in nature and are based on problem-specific characteristics and related intuitions. On the other hand, *metaheuristic search approaches* are problem-independent generic procedures. As such, metaheuristic approaches need to be adapted to the specific problem to be solved, and their parameters are to be tuned as well. In general, metaheuristic algorithms suffer from the lack of explicit exploitation of problem-specific characteristics. Note that metaheuristic algorithms typically adopt randomised selection strategies [63]. Thus, they are to be evaluated using statistical methods. Depending on the number of current solutions in each lap of the loops, search algorithms could be *single-solution* based and *multiple-solution* or *population* based. Many metaheuristic algorithms, such as *simulated annealing* and *iterated local search*, are single-solution based, while others, such as *genetic algorithms*, *memetic algorithms*, and *scatter search algorithms*, *differential*

evolution algorithms, *particle swarm optimisation algorithms* are population-based. Monte Carlo search techniques are one of the popular metaheuristic search methods in general and also in PSP. *Monte Carlo* search techniques depend on repeated random sampling and evaluation. Monte Carlo methods are simple, but because of an arbitrary amount of randomness could take a very long time to obtain good solutions [64]. Nevertheless, for metaheuristic algorithms in general, we refer to a few survey papers and textbooks [63, 65–69]. Metaheuristic approaches have been used heavily in lattice-based simplified PSP but considering the focus of this paper, we discuss them briefly in Section 24.

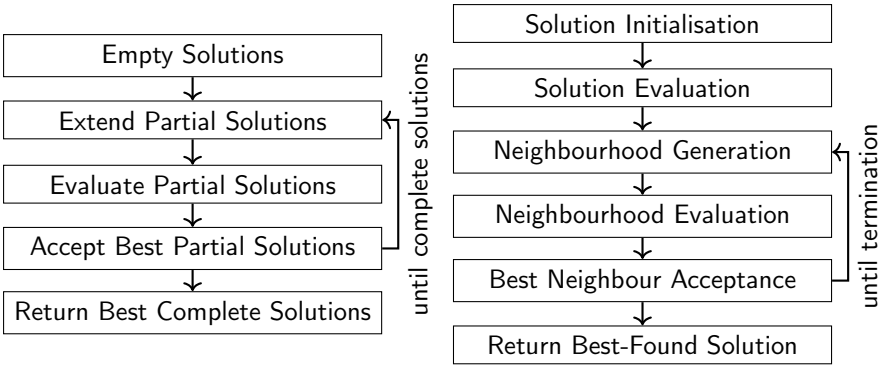

Fig. 6 Search-based optimisation frameworks: constructive search (left) and perturbative search (right)

4.3 Machine Learning Approaches

Machine learning approaches learn generalised input-output relationships from given training data using statistical analysis techniques and then predict output for given input test data. Because of the inherent nature of the generalisation process and the inevitable practical reasons for not having enough data that could exhaustively capture all possible scenarios, machine learning approaches typically can neither be perfect in prediction *accuracy* nor be complete in covering all possible cases. Nevertheless, recent machine learning approaches have made automated learning from training data possible almost without needing any explicit programming. So nowadays, machine learning approaches have been extensively used in solving many real-world computational problems [70]. Several traditional machine learning approaches include Support Vector Machine (SVM), decision tree, K-Nearest Neighbour Algorithm, and Neural Network. For details of such methods, we refer the reader to textbooks and survey papers [71–74]. However, we provide a brief overview of neural networks since research in the neural network-based system is dominating lately in machine learning in general and in PSP in particular.

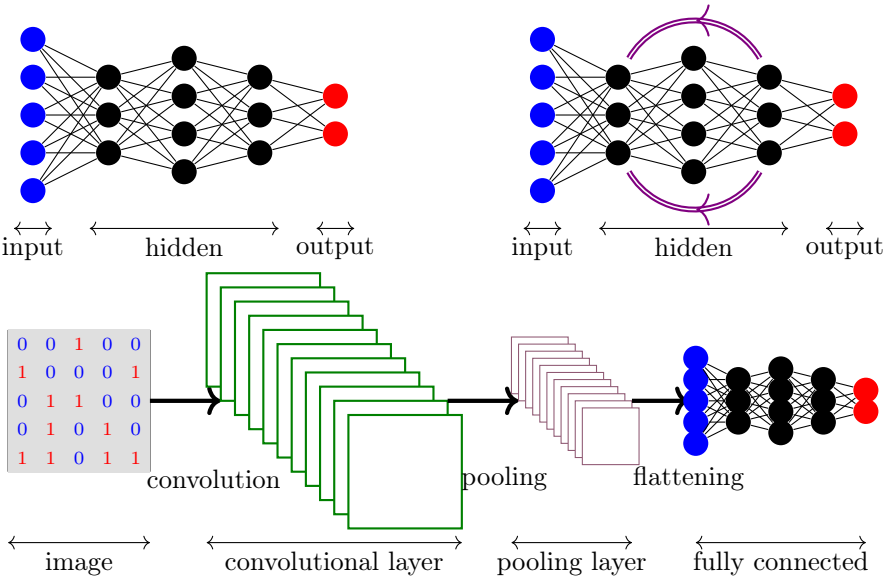

Fig. 7 Various neural networks: fully connected (top-left), recurrent (top-right), and convolutional (bottom)

Neural networks vaguely mimic interconnected neurons in a biological brain. The simplest example of a neural network is a *feed forward* neural work, in which artificial neurons are connected in layers, and data pass through the neural network from the input layer to the output layer. Fig. 7 (top-left) shows a fully connected neural network (FCNN) where each neuron in a preceding layer is connected to each neuron in the succeeding layer. A neural network is called a *deep neural network (DNN)* when it has more than one hidden layer; otherwise, it is a shallow neural work (SNN). Neural networks may have backward connections from later layers to the earlier layers. A recurrent neural network (RNN) is one such example. An RNN is shown in Fig. 7 (top-right). There exist other more complex types of neural networks that include convolutional neural network (CNN), bidirectional recurrent neural network (BRNN), long short term memory (LSTM), residual neural networks (ResNets), and generative adversarial networks (GAN). We refer the reader to textbooks and survey papers on DNNs [75–77]. Nevertheless, Fig. 7 (bottom) shows a convolutional neural network. Typically RNNs are used in learning sequences, while CNNs are used in image processing. DNNs, in general, need a huge volume of training data and huge computational effort in training.

5 Protein Structure Prediction Research

Search and optimisation problems typically have precisely defined constraints and objective functions. As such, the quality functions and ranking values of the solutions are well-defined. Even in machine learning algorithms in which,

from a user's point of view, the input-output relations are captured by the model, to an AI modeller, the search involved in finding the model has a precisely defined objective function (e.g. the loss function in neural networks) determining the quality of the model. So in AI, in general, the main research effort is spent on solution representation and search algorithm design where these two interact with each other. However, in PSP, one big challenge is the absence of a precisely defined energy function. Besides designing the sampling procedure in search, colossal research effort is needed in designing an actual or proxy energy function. Thus, the challenge for AI is compounded in PSP. However, it is somewhat surprising since there are well-defined geometry-based metrics to compare PSP solutions, and from AI perspectives, these metrics could be used as proxy energy functions, particularly in machine learning-based approaches, even if not in search-based optimisation algorithms. Fortunately, this has been realised well in recent times, and the end-to-end approach [54] has effectively shown the way, and AlphaFold2 [24] has then followed it.

PSP research has been largely driven by the Critical Assessment of protein Structure Prediction (CASP) [78] competition that has been taking place every two years since 1994. The competition has many tracks to assess the performance of computational methods for various aspects of PSP. We refer to the competition to get access to its excellent resource repository. Considering the PSP processing pipeline, we list several key research areas below. We later discuss them in more detail.

1. Protein Structure Representation: Develop effective representations of protein structures. Cartesian coordinates of the atoms are, of course, there. Moreover, ϕ-ψ or θ-τ based angular representations are currently used for main protein chains, while χ angles are used for side chains. Fig. 8 (left) shows protein main chain representation using dihedral angles ϕ, ψ, and ω. A dihedral angle can usually be defined using four points as shown in Figure 8 (right). However, each dihedral angle could also be represented by three planar angles [79]. So, dihedral angles ϕ, ψ, and τ could also be replaced by planar angles. Developing alternative representations is useful since one method could actually be used for structure representation while another could act as a constraint.

2. Developing Evaluation Metrics: Develop appropriate evaluation metrics to make a meaningful comparison, ranking, and validation of PSP-related experimental results. However, among all evaluation metrics, conformation comparison metrics are the key.
   (a) Machine Learning Accuracy: Precision, recall, mean absolute error (MAE), mean squared error (MSE) and other statistical methods are generally used. Develop further context-specific effective measures for the approaches used.

(b) Structure Accuracy Metrics: Design appropriate metrics to ascertain the quality of the whole or some part of a predicted protein structure compared to the protein's native structure. For example, root mean square distance (RMSD) is a standard metric.

3. Machine Learning Approaches: Designing input features and machine learning models is the main focus here. However, machine learning approaches could be used to predict final protein conformations as well as to learn intermediate constraints that could later be used as input features for further machine learning or in conformational refinement by search-based optimisation algorithms. We mainly discuss recent deep learning approaches for the main chains of proteins.
   (a) Input Feature Design: Input features could be solely from the given protein or other proteins in a protein library. Recent input feature design methods mostly rely on multiple sequence alignment (MSA) of the target protein and related proteins. The MSA sequences are then used directly as features or in the computation of more complex features.
   (b) Backbone Structure Prediction: For a given protein, predict its 3D structure using a suitable representation method, e.g. ϕ-ψ or θ-τ based methods. Protein backbone chains could have N, C^α, and C atoms or just C^α atoms.
   (c) Secondary Structure Prediction: For each residue in a protein, predict the 3- or 8-state secondary structure type of the residue or develop further classification categories. This is only about the secondary structure type of a residue but not the position of the secondary structure with respect to other secondary structures. Using this information, rigid secondary structures such as helices and sheets could be roughly constructed since rigid secondary structures have narrow ranges of ϕ and ψ angles. However, this does not really help construct the coil-type secondary structures as any angle value is possible.
   (d) Geometric Constraint Prediction: Predict various geometric constraints of a protein. The geometric constraints could be used as input features for further machine learning or could also be used in the refinement of somehow-already-predicted protein structures. Besides predicting the following constraints, further constraints could be explored.
      (i) Contact Map Prediction: For each pair of residues, predict whether designated atoms (e.g. C^α, C^β, or other atoms) of the residues are within a certain distance threshold (e.g. 8Å).
      (ii) Distance Map Prediction: For each pair of residues, predict the distances between designated atoms (e.g. C^α, C^β, or other atoms) of the residues. The distances could be represented by real values or by bins denoting distance ranges.
      (iii) Angle Map Prediction: For each pair of residues, predict various angular orientations (already defined or newer) of the residues. These

angles essentially help orient lines and planes comprising atoms with respect to other lines and planes.

(iv) Covalent Bond Prediction: Predict certain bonds such as disulphide or hydrogen bonds that capture various types of moderate or long-distance interactions between residues in a protein.

(e) Structure Accuracy Prediction: Given a decoy structure of a protein, predict the accuracy of the decoy structure using a given structure accuracy metric. Search-based optimisation approaches often provide several decoy structures selected as the best in terms of their energy or scoring functions. However, the best with respect to the energy or scoring function is not necessarily the best concerning the structure accuracy metric used. So a structure accuracy prediction method could be used to select one decoy structure. Structure accuracy prediction is also known as *estimation of model accuracy (EMA)*.

(f) Abstract Local Structure Prediction: For a given protein, predict its local structures at various abstraction levels. This is important since structures are more conserved than sequences [80]. For example, one pertinent question is as soon as a secondary structure is formed, whether the amino acid sequence in the secondary structure matters much in the overall folding of the 3D structure of the protein or whether the shape of the secondary structure becomes the key factor. Proteins naturally exhibit local structuring at various hierarchical levels. Secondary structures are towards the bottom of the hierarchy, but super secondary structures and secondary structural motifs are towards the middle. Hierarchical modelling is interesting from an AI perspective. Hierarchical approaches would help manage the scalability and complexity of huge proteins by apportioning the emphasis at various abstraction levels of the protein structure prediction process.

4. Search-Based Optimisation Approaches:

(a) Conformation Comparison Metrics: Design appropriate metrics for the conformations generated and evaluated during the search. Deep learning-based approaches somewhat implicitly do this via their loss functions and so do not need such metrics.

(i) Energy Function Design: Design energy functions at the all-atomic level where information only from the given protein is utilised. CHARMM [42] and AMBER [43] are such energy functions.

(ii) Scoring Function Design: Design scoring functions that can effectively act as proxies to the unknown energy function. Scoring functions usually are simplified at the amino acid or residue or any local structure or abstraction level. Recent scoring functions often use predicted contacts, distances, and angles between residue pairs.

(b) Structure Construction: Construct protein structures from the values of the variables in a given representation method. From ϕ-ψ or θ-τ angles,

protein backbones can be constructed, and from χ angles, side chains can be constructed.

(i) Random Construction: Construct a protein's structure entirely randomly since the structure will later undergo refinement. However, randomness could be controlled by exploiting various information, e.g. when the secondary structure types of the residues (using angle ranges for secondary structures), when predicted angle values are known with some error margins (from machine learning algorithms), or when probability distributions of certain angles are known (e.g. angle probability lists (APL) [81]). In all these cases, each angle value is selected randomly.

(ii) Fragment Assembly: Construct a protein's structure using the fragment assembly approach where a fragment is a small subsequence of a protein's amino acid sequence [82]. Fragments of the given protein could be searched in a fragment library, and the angles of the matching fragments could be used under probabilistic selection approaches. Essentially, fragments of a protein could overlap with each other. After constructing an initial conformation, the fragment assembly approach could also be used in refinement. So we do not discuss this separately.

(c) Structure Refinement: Given an already constructed protein structure, refine the structure, typically using a search algorithm to improve its quality in a given scoring function. The end target is, of course, to improve a given structure accuracy metric. There is a model refinement category in CASP to improve the quality of some of the best models produced by existing structure prediction methods. However, CASP allows any type of method, but here we mainly refer to the search-based methods.

(i) Refine using Energy Functions: Perform structure refinement search using energy functions such as CHARMM [42] or AMBER [43]. For a given protein, just exploit the information from the protein itself and its amino acid sequence and not from any other proteins or their structures. This is the ideal case of Anfinson's dogma.

(ii) Refine using Scoring Functions: Perform structure refinement search, e.g. by RaptorX [83] and trRosetta [20]. For a given protein, exploit information from the protein itself and other proteins, but those proteins should not be homolog (more than 30% sequence similarity) of the given protein.

5. Side Chain Prediction: For each residue in a protein with a backbone structure already constructed, predict the side chain angles by machine learning or search-based optimisation approaches to construct the whole structure of a protein.

6 Some Generic PSP Concepts

We discuss various generic concepts used in PSP research and refer to the concepts later as we provide detailed discussions on the subareas. This is to avoid repetition of describing the same concepts in various contexts.

1. Classification Prediction: Classes are usually represented by using *discrete values*, and more importantly, classes are *orthogonal* to each other. However, classes could be represented by probabilities as well. For classification problems, e.g. in secondary structure prediction, the class with the highest probability is usually taken as the output class.
2. Real Value Prediction: Real values, e.g. angles or distances, are floating-point numbers from specific ranges. Prediction of real values is considered as a regression problem in machine learning. One key difficulty with real value prediction, mainly when the possible range is quite extensive, is balancing the relative errors for small and large values. To deal with this difficulty, sometimes an extensive range of real values is split into several small ranges or *bins* of values and then a classification problem is solved rather than the typical regression problem. However, the difficulty with the binned approaches is that the bins are not orthogonal. So the prediction of the succeeding or preceding bins does not necessarily mean an utterly wrong prediction, which is the case in classification problems. The binned approach is also valuable for search-based algorithms when a range is discretised into bins, and during the search, only one value from each bin is considered.
3. Long-Range Interaction: Usually $|i - j|$ is defined as the *sequence separation* between the residues with indexes i and j. Because of the folding of the main chain of a protein, residues having long sequence separation could come physically close to each other. Since structures are more conserved than sequences [80], residues that are in physical proximity exhibit coevolutionary characteristics. So the interactions between residues having long sequence separation but within physical proximity are interesting. So by long-range interaction, we usually mean close proximity of residues having long sequence separation.
4. Using Sliding Windows: In order to capture the local context of the residues in a protein, the *sliding window* concept is often used. A window is defined with several residues before and the same number of residues after for a given residue. Since a window is defined for each residue, windows associated with successive residues differ by one residue on either side. Hence, the windowing method is called the sliding window method. Sliding windows are suitable for capturing short-range interactions but not good for long-range interactions. Moreover, the length of the sliding windows affects the performance much.
5. Using Entire Proteins: To capture long-range interactions among residues, all residues in a protein are typically used at the same time. However, the use

of the entire proteins could impose an upper limit on the size of the solution model, particularly on the deep learning architectures. Moreover, in such a case, shorter proteins need padding with do not care values in the input features, which essentially means introducing noise in the model for the shorter proteins. Furthermore, the effect of very long-distance interactions among residues is not apparent since long distances result in negligible energy values as per physics-based energy functions. An alternative could be to chop a protein into multiple (perhaps overlapping) pieces and work with the pieces.

6. Residue Level Prediction: In machine learning algorithms for proteins, predictions can be made at the residue level, e.g. for one residue or one residue pair at a time. In such a case, the sliding windows method might capture the local context of the residue or residue pair. However, to capture long-range interactions, residue-level input-output might not be effective. In such a case, all residues in the entire protein or a piece of protein are often used simultaneously.
7. Single Sequence-Based Methods: Current deep learning methods for proteins often involve building an MSA profile for a protein sequence library. These MSA profiles are then used as input features. This process is somewhat contrary to the basic hypothesis that a protein's amino acid sequence alone determines its 3D native structure. So, there is a body of machine learning research that does not use MSA features and instead extracts information only from the given protein. Such methods are called *single sequence-based methods*. However, these methods are still not based on the aforesaid basic hypothesis since any machine learning approach in PSP strives to learn the general underlying characteristics among the training proteins. So essentially, their performance for a given protein depends on the information captured from other training proteins.
8. End-to-End Approaches: Starting from the input, delivering the final output without needing any intermediate step is called an *end-to-end* approach. In PSP, deep learning or search-based optimisation algorithms often minimise certain loss functions or scoring functions, which are not necessarily structured accuracy metrics such as RMSD, TM-score, or global distance test (GDT) scores. If the final performance is measured in one metric, but the prediction method uses another metric in its optimisation process, then arguably, one would not get the best results in terms of the final performance metric. An end-to-end approach strives to obtain protein structures starting from amino acid sequences with an explicit focus on minimising a structure accuracy metric.

7 Protein Structure Representation

From AI perspective, solution representation is a key to the performance of a solving algorithm. Solution representation directly determines the size of

the solution space. Arguably, the fewer the variables, the smaller the search space. Note that there is a body of literature on lattice-based simplified protein representations that attempt to reduce the search space, but considering the focus of this paper, we only very briefly discuss them in Section 24.

7.1 Existing Representation Methods

Protein structures could easily be represented by using 3D Cartesian coordinates of the atoms. This representation is used at the PDB format [6]. However, such a representation is often not suitable for computational approaches for PSP. This is because such representation allows infinite numbers of translational and rotational symmetries, and the degree of freedom of the search space becomes enormous. So in computational approaches for PSP, protein structures are rather typically represented using sequences of local coordinate systems based on dihedral or planar angles and standard bond distances. Fig. 8 shows that protein's main chains can be represented using dihedral angles ϕ, ψ, and ω. A dihedral angle can be defined using four points, as shown in the right part of Figure 8. As such each of the angles ϕ_i, ψ_i, ω_i for a residue with index i is defined respectively by each four consecutive atoms from the sequence C_{i-1}, N_i, C^{α}_i, C_i, N_{i+1}, C^{α}_{i+1}. However, for most proteins, ω angles are fixed at 180° since peptide bonds are planar and rigid and hence do not allow free rotation around them. Note that protein backbones can also be represented by using planar angles θ_i defined by the atoms $C^{\alpha}_{i-1}, C^{\alpha}_i, C^{\alpha}_{i+1}$ and dihedral angles τ_i defined by the atoms $C^{\alpha}_{i-1}, C^{\alpha}_i, C^{\alpha}_{i+1}, C^{\alpha}_{i+2}$. This representation does not include N, C, and any other atoms. Protein side chains can be represented by using dihedral angles such as $\chi_1, \chi_2, \ldots$, but for brevity, we do not show their exact definitions in this paper. Note that construction of main chains from backbone angles is called *folding* since the results are the overall folded shapes, while construction of side chains from side-chain angles is called *packing*.

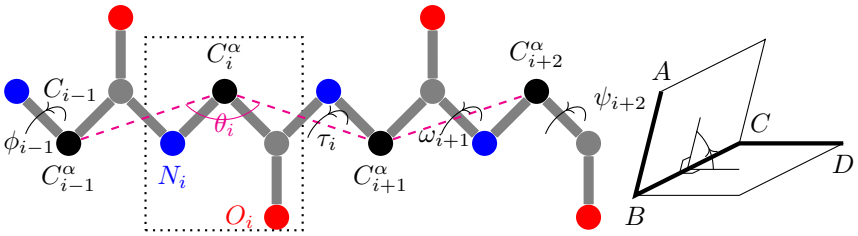

Fig. 8 Backbone angles of a protein structure (left) and a dihedral angle from four points (right)

7.2 Search Space Exploration

Considering any real-values are possible in $[-180°, +180°]$ for the planar or dihedral angles, in either of ϕ-ψ or θ-τ based backbone representations, the

conformational search space is clearly astronomical in size. Any χ angles in the side chains can also take any values from $[-180°, +180°]$, and the related search space is huge, too. However, as per Levinthal's paradox [84], proteins can still naturally fold into their native 3D structures within nanoseconds. Does nature explore the entire search space to find the native structure? A common understanding in protein structure prediction is that perhaps each protein obtains its stable native structure when its free energy is minimal. However, no such energy function is precisely known so far. As such, exploring scoring functions that can potentially be used as the surrogate models of the underlying energy function is also a great challenge in computational PSP. From the computational perspective, the purpose of a scoring function or an energy function is to provide a means to determine the better conformation among any given two conformations of the same protein.

7.3 Learning and Searching

Exploring PSP conformational space effectively needs both machine learning and search-based optimisation approaches. Search-based optimisation approaches strive to find optimal solutions by performing a search on the solution space of a given problem instance. Machine learning algorithms perform a search on the space of possible learning models such that those models capture the input-output mapping from available solved instances of the problem. Once a learning model is obtained, machine learning algorithms do not need any further search to find the solution to any given future problem instance. However, search-based optimisation algorithms need to search for each given problem instance. For PSP, amino acid sequences are problems, and native conformations are solutions. So machine learning algorithms could predict ϕ, ψ, θ, τ, and χ angle values so that the protein structures could be *constructed* directly. On the other hand, search-based optimisation algorithms could explore the conformational space and perform continual *refinement* starting from scratch (i.e. from an empty confirmation) or a randomly generated confirmation. Moreover, search-based optimisation algorithms could also start from conformations constructed from the predictions made by machine learning algorithms and perform continual refinement of the conformations by exploiting constraints learned by different machine learning algorithms. The interplay between searching and learning and the intertwined characteristics make PSP a fascinating problem from AI perspective.

8 Structure Accuracy Metrics

In AI methods, solution quality is usually clearly defined in terms of the value of the objective function. However, there is no such clearly defined objective function in PSP problems. As such, research effort is needed to define accuracy metrics that could effectively compare a predicted structure of a protein with its native structure and give a score. The research problem can rather be stated

as "given two 3D structures of the same protein, measure their structural difference".

8.1 Existing Accuracy Metrics

We refer the reader to the CASP website for a list of existing structure accuracy metrics. We have already mentioned RMSD. Besides RMSD, template modeling score (TM-score) [85], GDT score [86], and local distance difference test (lDDT) score [87] are often used to compare the whole or some parts of the potential decoy structures of the same protein.

Note that RMSD computes the average deviation of the atoms of a given structure from the corresponding atoms of the native structure of the same protein. As such, RMSD requires the superposition of the given structure on the native structure. The Kabsch algorithm [88] is used to compute the optimal rotation matrix to align corresponding atoms of two conformations and thus to get the superposition. However, RMSD could be greatly affected by outliers in local structures and also by the length of the proteins. Furthermore, RMSD can be dominated by divergent loops that can obscure local regions of similarity. A distance version of RMSD (dRMSD) is superposition free and based on the distances between atom pairs in the given structure and the distances between corresponding atom pairs in the native structure.

Unlike RMSD, TM-score is length independent, limiting the impact of divergent pairs of atoms in superimposed structures. In TM-score, different subsets of atoms are superimposed rather than superimposing the entire structures. In this way, an approximately optimal superposition can be obtained by sampling a large number of subsets. Because of the numerous local superpositions, the TM-score algorithm is slower than the RMSD calculation. The TM-score usually ranges from 0 to 1, with 0 indicating no match and 1 indicating a perfect match between two structures. According to stringent statistics of PDB structures, scores less than 0.17 correspond to randomly picked unrelated proteins, whilst scores more than 0.5 presume the same fold in native-like structure [85]. Overall, the TM-score weights minor distance errors more strongly than larger distance errors and so is more sensitive to the global structural similarity than to the local structural errors.

GDT score is one of the most accepted accuracy metrics used in measuring the quality of predicted protein structures. Given a set of distance thresholds (e.g. 1Å, 2Å, 4Å, and 8Å) [89, 90], GDT maximises the percentage of superimposed residue pairs under each threshold and computes the average of the percentages. The GDT score ranges from 0% to 100%. The closer the GDT to 100%, the more accurate the model's backbone, with values above 80% indicating that local and global features are most accurately predicted and values below 20% indicating mostly random models. Computing the GDT score for a given predicted structure is conjectured to be NP-Hard. So, available GDT computing methods are heuristic-based and might not give the optimal scores.

lDDT is superposition-free like dRMSD and evaluates the local distance differences of all atoms in a structure. lDDT adopts a similar approach to

GDT in computing percentages of atom pairs having their distances predicted within certain threshold tolerance levels (e.g. 0.5Å, 1Å, 2Å and 4Å) and then taking the average of the percentages.

8.2 Further Research Challenges

As already noted, RMSD, TM-score, and GDT are orientation-dependent as they need superposition. Finding a superposition of one structure on another is itself an optimisation problem. So a search procedure is involved, and advanced AI techniques could, of course, play a significant role. Basically, each scoring function has its own pros and cons in terms of how it measures global and local similarities between the two structures. So further research is needed to obtain a more meaningful comparison of protein structures.

9 Input Feature Design

Input features are essential for machine learning approaches. In PSP, the primary input features are the amino acid residues and the amino acid types in the amino acid sequence. Next, the properties of entire amino acid sequences could also be among the primary input features. Input features could be extracted from other amino acid sequences that are in some way similar to the given protein's amino acid sequence. Moreover, some predicted features could also be used as further input features for subsequent prediction models.

9.1 Single Sequence Features

Besides amino acid sequences of the proteins, various other characteristics of the amino acids and the proteins have been used as input features of the machine learning algorithms. For example, 7 physicochemical properties (7PCP) such as steric parameter (graph shape index), hydrophobicity, volume, polarisability, isoelectric point, helix probability, and sheet probability [91] have been used in backbone angle prediction [92–96] and secondary structure prediction [93–95]. Also, solvent accessible surface area (SASA)[97] [97] has been used in backbone angle prediction [92]. For a given protein, the higher the quality of the structures, the fewer the amino acids in contact with the solvent [98]; this is somewhat related to the hydrophobicity property. Moreover, Shannon entropy has been used in contact map prediction [12] and distance map prediction [99]. Shannon entropy measures the information content of the amino acid sequences of the proteins.

9.2 MSA Based Features

In the course of time, for a given protein, by using MSA, information has been extracted from other proteins that have similar amino acid sequences. For example, position specific scoring matrix (PSSM) generated by PSI-BLAST [100] has been used in backbone angle prediction [92–96], secondary structure prediction [94, 95, 101], contact map prediction [14, 102], and distance

map prediction [20, 99]. PSSM describes observed frequencies of various amino acids at every residue position with respect to a given sequence library. PSSM essentially tries to capture somewhat weak similarities between amino acid sequences of various protein families [103]. Further, Hidden Markov Model (HMM) profiles produced by HHBlits [104] have been used in backbone angle prediction [95, 96, 105], secondary structure prediction [94, 95, 105], and contact map prediction [14]. HMM can capture various hidden properties, e.g. secondary structural information [106], from the amino acid sequences. Furthermore, covariance matrix has been used in contact map prediction [102, 107, 108] and distance map prediction [99].

9.3 Coevolutionary Features

PSSM or HMM does not encode any spatial information about the residues in the protein structures. Later, contact-based input features have been designed [109, 110] to capture spatial information based on the knowledge that residues that are in close proximity coevolve with each other [111]. However, coevolution is sometimes problematic when residues A and C are not in contact but appear to be coevolutionary with each other since they both are separately coevolutionary with respect to B. Nevertheless, while HMM can model very short-range residue correlation, Markov random field (MRF) can model long-range residue correlation and can also represent protein families. MRF has been used as summarized pairwise potentials [112, 113] or in raw format [114, 115]. CCMPred has been used in contact map prediction [14, 102, 116–119]. Precision matrix has been used in contact map prediction [108, 120] and distance map prediction [99]. The pseudolikelihood maximization matrix has been used in contact map prediction [120]. Compressed covariance-matrix has been used in contact map prediction [121]. A reduced precision matrix has been used in contact map prediction [108, 121]. FreeContact has been used in contact map prediction [102, 119, 122]. Contact potential has been used in contact map prediction [12]. Contact density has been used in distance map prediction [99, 123, 124].

9.4 Raw MSA as Features

Many MSA-based coevolutionary features are costlier both in time and memory. So, instead of computing those features from MSA, raw MAS has instead been used very recently, first with mixed results [125] and then also with promising results [24].

9.5 Predicted Values as Features

Some characteristics predicted by one machine learning algorithm have also been used as input features of another machine learning algorithm. For example, predicted secondary structures had been used in backbone angle prediction [126]. Also, predicted contact maps were used in backbone angle prediction [95] and secondary structure prediction [95].

9.6 Further Research Challenges

Overall, MSA, covariance, or coevolutionary features contain crucial information about protein structures. They have also led to recent great progress. However, they are far away from the primary features based on amino acids and amino acid sequences. Moreover, these features depend on given protein sequence libraries. So their performance (also that of machine learning) is restricted by the sequence libraries used. Recently AlphaFold2 [24] addressed this with success but by using almost all known and unknown proteins and using enormous computational resources. This clearly raises concerns from AI perspective. It is well-known that the more data exploitation, the better the accuracy. Also, it is well-known that the more time spent, the better the accuracy. However, AI strives to obtain high-quality results using as little data and less time as possible. So effective and compact features are sought.

10 Backbone Angle Prediction

Protein backbone or main chain angle prediction deals only with the backbone chain of the proteins leaving their side chains out. Protein backbone angle prediction is essential since using ϕ-ψ or θ-τ values, a protein's backbone chain can be precisely constructed. We describe machine learning approaches mainly here, while search algorithms are described later in separate sections.

10.1 Existing Backbone Angle Prediction Methods

Recent backbone angle prediction methods include SPIDER [92], SPIDER2 [127], SPIDER3 [94], RaptorX-Angle [128], DeepRIN [129], SPOT-1D [95], NetSurfP-2.0 [105], CRNN [130], OPUS-TASS [131], and SAP [126].

Most early backbone angle prediction methods predict backbone angles on a residue basis. However, the most recent methods simultaneously predict all residues in a protein to capture global interaction between residues.

Recent backbone angle prediction methods are mainly DNN based. For example, they use stacked sparse auto-encoder DNNs [92], successive iterated DNNs [93], RNNs [130], CNNs [105, 128], LSTM [105], LSTM-BRNNs [94, 95, 131], and ResNet [95, 128], deep residual inception networks (DRIN) [129] and ensembles of DNNs [95, 131].

Recent single-sequence based backbone angle prediction methods include SPIDER3-Single [132], SPOT-1D-Single [133], and ProteinUnet [134]. SPIDER2-Single uses LSTM-BRNNs, SPOT-1D-Single uses LSTM-BRNNs and ResNets, and ProteinUnet uses Unet CNN.

Input features used in these methods include PSSM [92–96], 7PCP [92–96], ASA [92], HMM profiles [95, 96, 105], contact maps [95], and PSP19 [131]. Beside CASP datasets, other datasets used in evaluating these methods include PISCES [135], SPOT-1D [95], PDB150 [129], and CAMEO93 [136].

The first machine-learning method SPIDER [92] makes sequence-based prediction of θ and τ angles with MAE of 9° for θ and 34° for θ. SPIDER2 [127]

Table 2 Performance comparison of recent backbone angle prediction methods in terms of MAE in degrees. In the table, KMC denote K means clustering algorithm, PSFM denotes position-specific frequency matrix, EP denotes evolutionary profiles, STP denotes state transition probabilities, and LAD denotes local alignment diversity.

Method	Year	Features	Dataset	Techniques	ϕ°	ψ°	θ°	τ°
SPIDER [92]	2014	PSSM, 7PCP, SS, SASA	PISCES, TR4590, TS1199	DNN	22	33	9	34
SPIDER2 [127]	2017	PSSM, 7PCP	TR4590, TS1199, CASP11	DNN	19-21	30	8	32
SPIDER3 [94]	2017	PSSM, 7PCP, 30-D HMM	TR4590, TS1199, TS115	LSTM, BRNN	18-21	27	8	30
RaptorX-Angle [128]	2018	PSSM, PSFM, SASA, SS	PDB25, TS1267, CASP11, CASP12	KMC, CNN, ResNet	18-20	27-33	-	-
DeepRIN [129]	2018	PCP, PSSM, HMM, SS8	CullPDB, CASP10, CASP11, CASP12	DRIN	17-20	26-31	-	-
SPOT-1D [95]	2018	PSSM, HMM, 7PCP, SPOT-Contact	PISCES [137], TEST2018, TEST2016	LSTM, BRNN, ResNet, Ensemble	17	25	7	25
NetSurfP-2.0 [105]	2019	HMM, STP, LAD	CASP12, TS115, CB513	CNN BRNN, LSTM	17-20	26-32	-	-
CRNN [130]	2020	HSSP	CullPDB, CB513, CASP10, CASP11, CASP12	Clustering RNN	17-19	25-30	-	-
OPUS-TASS [131]	2020	PSSM, HHM, 7PCP, PSP19 ([138, 139])	SPOT-1D, CAMEO93, TEST2016, TEST2018, CASP12, CASP13, CASP-FM	CNN, LSTM, Trans-former	16	22-24	-	-
SAP [126]	2020	PSSM, SS8, 7PCP, ASA	SPOT-1D, TEST2018, TEST2016	FCNN	16	19	6	21
SAP4SS [140]	2022	SS8, PSSM, 7PCP, HMM	SPOT-1D, PDB150, CAMEO93	FCNN	15.6	18.9	6.0	21.7
IGPRED-MultiTask [141]	2022	PSSM, 7PCP, struc-tural profile	OPUS-TASS, CASP12, CASP13, CAMEO93	CNN, mGCN	15	18	-	-

is an improved version of SPIDER with three hidden neural network layers. Next, SPIDER3 uses BRNN with LSTM to contemplate the multiple features in addition to the above properties for predicting secondary structure. SPIDER3 reached up to 19°, 30°, 8° and 30° MAE values ϕ, ψ, θ, and τ angles [94].

RaptorX-Angle [128] predicts real-valued angles by blending deep learning and clustering techniques. RaptorX-Angle was tested on a subset of PDB25 and the targets in CASP. RaptorX-Angle outperformed SPIDER2 [127] in terms of MAE and Pearson Correlation Coefficient (PCC). Another study [128] presented an unconventional and more precise prediction of dihedral angles, which may accelerate the studies on protein structure prediction. DeepRIN [129], is another method for the estimation of $\phi - \psi$ bond angles. The input to DeepRIN is a feature matrix demonstrating the physicochemical properties of protein and amino acids, PSSM, HMM [142] and predicted 8-state SS. DeepRIN is created based on ResNets and inception networks. Overall, DeepRIN outperformed the best state-of-the-art tools like SPIDER3 [94] significantly. In such cases, on average, DeepRIN decreased the prediction errors of ϕ and ψ angle by 2° and 5°, respectively [129].

SPOT-1D [95] uses ResNets and LSTM Cells BRNNs to predict backbone angles (ϕ, ψ, θ, and τ). NetSurfP-2.0 [105] employs LSTM in BRNNs to predict ϕ and ψ angles. Simpler Angle Predictor (SAP) is a recently developed deep learning-based method by [126] to improve protein backbone angle prediction. The study results revealed that the SAP could considerably outperform existing angle prediction methods: the variations are 6–8 concerning the MAE for some types of angles. A similar model of SAP called SAP4SS [140] has outperformed the existing state-of-the-art methods in terms of MAE for all four types of angles, which are from 1.5 to 4.1% compared to the best-known results. Clustering Recurrent Neural Network (CRNN) [130] is another modern deep learning method used further to advance the performance of protein structure prediction. In this approach, the entire protein sequence dataset was separated into multiple cluster subtrees. An RNN is trained for each cluster in the subtrees to learn the computationally easier local sequence-to-structure connection. After which, a CRNN was constructed to predict mainly torsion angles and secondary structures for backbone $C\alpha$ atoms of protein sequences. OPUS-TASS [131] proposes a protein backbone angles and secondary structure predictor based on CNN layers, LSTM layers and modified transformer layers. An ensemble of neural networks adopted here further improved the performance. In comparison, OPUS-TASS achieves the MEA 16.28° and 21.98° for final ϕ and ψ predictions, respectively.

In early approaches, the torsion angles were predicted in a few discrete bins though the latest techniques have engaged in predicting dihedral angles in real continuous values. However, real value prediction cannot give proper information on the probabilities of predicted angles. A study in [143] suggests a method to predict bond angles in fine grids of 5° by utilizing deep learning. According to these findings, the grid-based method can generate about 2–6%

higher accuracy in predicting angles in the same 5° bin than many present methods. The researchers further investigated the effectiveness of predicted probabilities at provided angle bins in discrimination of essentially disorder regions and in selecting protein models.

The state-of-the-art results in terms of MAE values for backbone angle prediction methods are about 15° − 16°, 18° − 19°, 6° − 7°, and 21° − 22° for ϕ, ψ, θ, and τ, respectively. The state-of-the-art MAE values for single-sequence based backbone angle prediction are roughly about 34°, 22°, 9°, and 34° respectively for ϕ, ψ, θ, and τ. So, there is significant room to improve the prediction accuracy. Table 2 shows the performance comparison of recent backbone angle prediction methods with a description of input features, dataset, technique used and MAE values (ϕ, ψ, θ and τ).

10.2 Further Research Challenges

In protein backbone angle prediction, to capture local interactions among residues, fixed-length sliding windows have been used while feeding the input features to the DNNs [92–94, 126] at the residue level. Determining the window size is a challenge. Moreover, windows cannot capture global interactions between residue pairs. So to capture non-local or long-range interactions among residues, entire protein sequences have been used as features [93, 94, 96, 132] or CNNs [105, 131] or LSTM-BRNNs [94, 95] and all angles for a protein have been predicted at the same time. Using entire protein sequences creates trouble with proteins having varying lengths of proteins. So fixed-length pieces of proteins could be used, but they have not been tried so far.

From the results of the state-of-the-art backbone angle prediction methods, prediction of the ϕ and ψ angles appears to be harder for coils than for sheets and to be the easiest for helices [126]. The rigid structures of helices and sheets and the flexible structures of the coils help explain these results. So one challenge in protein backbone angle prediction remains in achieving more accurate predictions for the coil residues. Moreover, secondary structure boundary regions are also troublesome for backbone angle prediction since secondary structure prediction itself could be inaccurate in those regions. This means that the prediction of backbone angles for residues around the middle of the helices and sheets is more straightforward because of local consistency.

One key challenge in protein backbone angle prediction is to get the predicted angle values such that the structures constructed using the predicted values have excellent quality. Certainly, accurately determined backbone angles will lead to the exact native structures. However, protein structures constructed using even moderately inaccurately predicted angles might be far from native structures. In terms of the resultant structures, errors in the predicted angles in one part of the protein will have a cascading effect on the other parts. Interestingly, the opposite effect could also be observed: prediction errors in one part of the protein might be compensated by that in another part, and hence the overall shape of the protein structure might still be good. Consequently, just minimising errors in predicted angles using MAE or MSE might not be

sufficient. A solution to this is to use an end-to-end approach with the structure accuracy metrics incorporated within the loss function. However, the incorporation of the comparison between the native structures and the structures produced using predicted angles within the loss function of the DNN models is slightly challenging. The challenge is both in constructing the 3D structures within the training process and then in translating the differences between the native structures and the constructed structures to the back-propagation of errors in the angles. End-to-end approaches have been used recently by NEMO [144], HMS-Casper [54], DMPfold [145], rawMSA [125], AlphaFold2 [24, 146], and RoseTTAFold [23]. This is an important direction from AI perspective because it is a straightforward argument: if the end results are measured in one metric, there is no point in minimising a different metric at any stage.

One indirect way to incorporate protein structures within the deep learning model is to use predicted geometric constraints such as contact or distance maps as input features of a backbone angle predictor, e.g. SPOT-1D uses contact maps [95]. This could essentially eliminate some of the difficulties related to 3D structure construction within the loss function of a DNN. Another alternative and more indirect way is to use more informative features that capture the covariance strength of inter-residue positions and have been used in contact or distance, or angle map prediction. Such features include CCMPred [116], FreeContact [122], ShannonEntropy [12], Contact Potential [12], Precision Matrix [108, 120], Pseudolikelihood Maximization Matrix [120], Compressed Covariance-Matrix [121], and Reduced Precision Matrix [108, 121]. Further work in these directions is needed.

From the state-of-the-art results obtained by various backbone angle prediction methods, the planner angle θ prediction appears to be easier than that of the other three dihedral angles ϕ, ψ, and τ. This could mean replacing each dihedral angle with three planar angles; thus, working with a representation comprising only planar angles could be beneficial. However, in that case, errors will be there with all three planar angles, and the resultant error in the dihedral angle computation could still be higher. So far, no research has been found in this direction. Further research could elucidate these hypotheses.

11 Secondary Structure Prediction

Protein secondary structure prediction is vital because 3D structures are essentially composed of secondary structures. When proteins fold, the secondary structures are first formed and then packed into the 3D native structures. Secondary structures offer information about protein activity, functions, and relationships [147]. Protein secondary structure can be classified into either 3 or 8 states[61]. Predicted secondary structures can be used as input for 3D structure prediction methods [140, 148–153].

11.1 Existing Secondary Structure Prediction Methods

Recent secondary structure prediction methods include Porter 4.0 [154], SSpro/ACCpro 5 [155], DeepCNF [156], SSREDNs [157], SPIDER2 [127], SPIDER3 [94], MUFOLD-SS [158], SPOT-1D [95], NetSurfP-2.0 [105], PORTER 5.0 [159], BetaDL [160], CRNN [130], OPUS-TASS [131], IGPRED [101], NetSurfP-3.0 [34], and IGPRED-MultiTask [141]. Most early secondary structure prediction methods predict secondary structure types for one residue at a time using sliding windows. Most recent methods try to capture global interactions between residue pairs and so make predictions for all residues in a protein at the same time. Nevertheless, the state-of-the-art results of secondary structure prediction in terms of accuracy are about 87% for 3-state and 77% for 8-state prediction. Naturally, an 8-state prediction is harder than a 3-state prediction. Moreover, classes in the 8-state prediction have some overlapping in their definitions, making accurate prediction difficult. However, a ceiling in the secondary structure prediction accuracy has been estimated to be 88% [161]. Therefore, given the current state-of-the-art, there is still a significant gap in 8-state secondary structure prediction. Many successful methods like SPOT-1D [95], NetSurfP-2.0 [105], OPUS-TASS [131], SPOT-1D-LM [162], and NetSurfP-3.0 [34] combine secondary structure prediction and angle prediction techniques together to develop a multitask learning methods. Note that NetSurfP-3.0 [34] takes advantage of recent developments in pre-trained protein language models [163] to improve the run-time of its predecessor by nearly two orders of magnitude while exhibiting comparable prediction performance.

Protein secondary structure prediction methods focus mainly on neural network-based architecture to obtain better prediction performance. For example, they use successive iterated DNNs [93], RNNs [130, 157], BRNNs [154], CNNs [105, 141, 156, 164], graph CNN [101, 141], LSTM [105], LSTM-BRNNs [94, 95, 131], and ResNet [95], deep inception networks [158] and ensembles of DNNs [95, 131].

Input feature selection is also an important part of this area. Input features used in these methods include PSSM [94, 95, 101, 141], 7PCP [93–95], HMM profiles [94, 95, 105, 141], contact maps [95], and PSP19 [131]. SPOT-1D [95] combined the contact map predicted by SPOT-Contact [14] into its input features to enhance its performance. OPUS-TASS [131] introduced a new input feature (PSP19) based on local structures of residues derived from OPUS-DOSP [138]. Besides the datasets from CASP competitions, other datasets used in evaluating secondary structure prediction methods include PISCES [135], and SPOT-1D [95].

Recent single-sequence based secondary structure prediction methods include SPIDER3-Single [132], SPOT-1D-Single [133], and ProteinUnet [134]. SPIDER2-Single uses LSTM-BRNNs, SPOT-1D-Single uses LSTM-BRNNs and ResNets, and ProteinUnet uses Unet CNN. The state-of-the-art results achieved by these methods in terms of accuracy are 72% to 74%. These results are arguably worse than those discussed above since MSA-based coevolutionary input features are not used in these methods.

Note that when homologous proteins based on sequence similarity and sequence-based structural similarity are used in building sequence profiles of the proteins, SSPro8 [155] archives about 93% accuracy in secondary structure prediction. Table 3 shows the performance comparison of recent 3- and 8-state SS prediction methods with a description of input features, dataset, and techniques used.

11.2 Further Research Challenges

As shown in Table 4, secondary structure prediction gives us narrow ranges (about 20°) for some SS types. Because of these narrow-angle ranges, one could view protein backbone angle prediction as a classification problem via SS type prediction, although backbone angles are actually continuous-valued. However, the narrow-angle ranges are only for helices and sheets and not for coils. This allows comparatively easier construction of helices and sheets once the constituent residues are known. Coil type secondary structures such as C, S, and T in Table 4 still have the full ranges. Moreover, about 40% residues in average proteins are coils [165]. Overall, secondary structures are a coarse-grained description of protein local structures where the angle ranges are somewhat arbitrarily defined based on ϕ-ψ distributions in Ramachandran plot [166]. Still, coils essentially have no well-defined structures. So solving protein secondary structure prediction does not necessarily make solving the protein structure prediction trivial.

The main difficulty in secondary structure prediction remains to make predictions in the boundary regions where one type of secondary structure ends, and another type begins. Prediction accuracy usually is very high for residues that are towards the middle of a secondary structure. Residues at the boundary regions often get incorrect secondary structure assignments by predictors [80].

Protein secondary structures essentially capture various areas in the Ramachandran plot. The 3-state or 8-state classifications are just different ways to capture various areas in that plot. In this context, based on narrower ϕ and ψ angle ranges, a 27-state classification has been proposed [167] recently to allow a more exemplary mapping of the ϕ-ψ space. While the 8-state prediction is more complex than the 3-state prediction, the 27-state prediction will be even more complicated. Nevertheless, no 27-state prediction method has been proposed, which could be a potential research direction. This might need more data per class and thus pose some challenges.

12 Contact Map Prediction

Contact maps are two-dimensional symmetric Boolean arrays representing contacts between residue pairs. Two residues in a protein are in *contact* if their distance in the native structure of the protein is no larger than a given threshold (typically 8Å). In this case, each residue in the distance computation is typically represented by the C^β atom (C^α for Glycine). Contact maps are

Table 3 3- (SS3) and 8-state (SS8) SS prediction accuracy (%). CDRP denotes context-dependent pseudo-potentials, GRU denotes gated recurrent units, mGCN denotes multigraph convolutional network, STP denotes state transition probabilities, and LAD denotes local alignment diversity.

Methods	Year	Features	Dataset	Techniques	SS3	SS8
SSpro ACCpro [155]	2014	MSA and profile probabilities	pdb_full, pdb_ante, pdb_post, TRAIN, RS126, EVA	BRNN	88 - 93	86 - 88
DeepCNF [156]	2016	PSSM	CullPDB, CB513, CASP10, CASP11, CAMEO	DNN	82 - 85	68 - 72
SSREDNs [157]	2017	PSSM, SA	CullPDB, CB513	deep CNN, GRU	83 - 84	66 - 73
SPIDER2 [127]	2017	PSSM, 7PCP	TR4590, TS1199, CASP11	DNN	81 - 82	-
SPIDER3 [94]	2017	PSSM, 7PCP, HMM	TR4590, TS1199, TS115	LSTM, BRNN	84	-
MUFOLD-SS [158]	2018	7PCP, PSSM, HMM	CullPDB, CB513, CASP10, CASP11, CASP12	DNN	83 - 86	71 - 76
SPOT-1D [95]	2018	PSSM, HMM, 7PCP, SPOT-Contact	PISCES [137], TEST2016, TEST2018	Ensemble, LSTM, BRNN, ResNet	87	77
NetSurfP-2.0 [105]	2019	HMM, STP, LAD	CASP12, TS115, CB513	CNN, LSTM	82 - 86	71 - 75
PORTER 5.0 [159]	2019	PSSM	PDB (2014 - 2017)	2-stage ensemble, BRNN, CNN	84	73
CRNN [130]	2020	HSSP	CullPDB, CB513, CASP10, CASP11, CASP12	Clustering RNN, K-means	84 - 87	-
OPUS-TASS [131]	2020	PSSM, HHM, 7PCP, PSP19 ([138, 139])	CAMEO93, TEST2016, TEST2018 CASP12, CASP13, CASP_FM	CNN, LSTM, Transformer	88 -89	77 - 79
IGPRED [101]	2021	PSSM, 7PCP, structural profile	CullPDB, EVAset, CASP10, CASP11, CASP12	Ensemble, CNN, mGCN	86 - 89	-
NetSurfP-3.0 [34]	2022	HMM, STP, LAD, ESM-1b [163]	TS115, CB513, CASP12, CASP14_FM	CNN, biLSTM,	79 - 86	70 - 75
IGPRED-MultiTask [141]	2022	PSSM, 7PCP, structural profile	OPUS-TASS, CASP12, CASP13, CAMEO93	CNN, mGCN	86 - 89	-

Table 4 Typical angle ranges for various secondary structure types [61].

SS	B	C	E	G	H	I	S	T
ϕ Range	[-130,-110]	[-180,+180]	[-130,-110]	[-59,-39]	[-67,-47]	[-67,-47]	[-180,+180]	[-180,+180]
ψ Range	[110,130]	[-180,+180]	[110,130]	[-36,-16]	[-57,-37]	[-80,-60]	[-180,+180]	[-180,+180]

translation and rotation invariant two-dimensional projections of 3D protein structures. Contact maps provide coevolutionary and functional information as well as spatial information for residue pairs [168, 169]. Fig. 9 shows the actual contact map of the protein with PDB ID 5Z02 using the native structure.

Protein contact can be classified into four classes based on how far separated the two residues are in the given protein primary sequence: local (apart from less than six residues in the sequence), short-range (separated by 6 to 11), medium-range (separated by 12 to 23 residues), and long-range (24+ sequence separation) [170]. While local contacts in a sequence capture secondary structure information, medium-range and long-range contacts are more important for protein structure construction accurately, but Long-range contacts are challenging to predict. Most contact map prediction approaches evaluate long-range contacts individually as they are the most meaningful of the three and the hardest to predict. Therefore, the primary focus of contact map prediction is to anticipate long-range contacts as accurately as possible.

All contacts are not equally useful for building 3D models. Note that predicted contact maps could often have probabilistic values instead of Boolean values denoting potential contacts. Nevertheless, predicted contact maps could be used as geometric constraints during the construction or refinement of the 3D structures of a protein by using optimisation-based search algorithms [171]. Predicted contact maps could also be used as input features of deep learning-based predictors for other aspects, e.g., in protein backbone angle prediction [95] and the EMA [172].

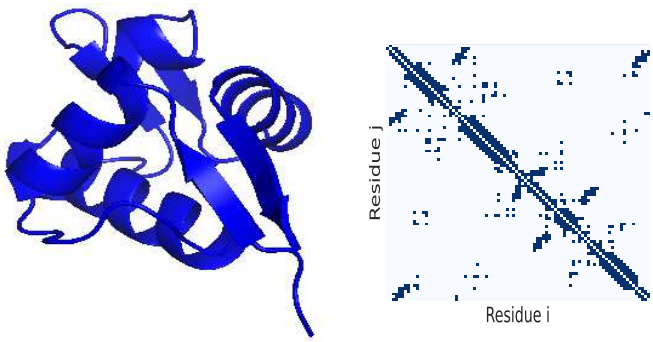

Fig. 9 The three-dimensional native structure of a protein 5Z02 (left) and its corresponding residue-residue contact map (right).

12.1 Existing Contact Map Prediction Methods

Predicted contacts are assessed using mainly precision, i.e., the number of correct contacts out of all predicted contacts in the given protein sequence. For many proteins, at least 8% of native contacts are adequate to rebuild the proteins fold. [170, 173]. Further, all proteins do not have their specific number of contacts proportional to the protein sequence length. Thus, it is typical to assess the top $L/2$, or just the top $L/5$ predicted contacts using precision, with L being the sequence length of the protein [13, 174]. The CASP contest recently concentrates on assessing predicted long-range contacts since short/medium-range contacts are relatively easier to predict (especially for proteins having β-sheets).

Recent contact map prediction algorithms include DNcon [175], DeepConPred [176], DeepCov [107], PconsC4 [177], RaptorX-Contact [13], DeepContact [117], DeepConPred2 [118], DNCON2 [119], DEEPCON [102], SPOT-Contact [14], DeepCDpred [178], ResPRE [108], MapPred [121] and TripletRes [15].

Adhikari and Cheng [179] summarize the machine learning-based contact map prediction methods. Xie et al. [180] explain machine learning-based methods and neural network-based techniques for contact map prediction. After deep learning-based protein contact prediction was first presented in 2012 [181, 182], various types of deep learning techniques have been developed to combine traditional protein sequence features with residue-residue co-evolutionary scores to substantially enhance the contact prediction accuracy [14, 102, 107, 119, 178].

DNcon [175] was the first successful approach to employ Deep Learning, and it outperformed all other competitors in the contact prediction category of CASP11 competition. RaptorX-contact [13] method employed deep layers consisting of two major ResNet modules and was the first method to use convolutional ResNets to predict contact maps for entire proteins. RaptorX-contact is capable of capturing very complex sequence-contact relationships and high-order contact correlations. RaptorX-Contact had better outcomes in CASP11 targets compared to other existent methods such as MetaPSICOV [12] and CCMPred [116]. RaptorX-Contact was also rated top in CASP12 and CASP13 under the contact prediction class.

ResPRE [108] predicts residue-residue contacts based on the MSA and the covariance matrix. The covariance matrix used in this method is converted into a precision matrix and then fed to the residual CNN. A novel idea of using a precision matrix gives value to this method. MapPred [121] utilises a metagenome sequence in a ResNet architecture. The use of the metagenome sequence is statistically significant. In addition to contact maps, MapPred also predicts distance maps and distance distribution. In DEEPCON [102], authors focused on improving the precision of medium-range and long-range contact. The neural network architectures examined in this work with alternating dilations and dropout predict contacts with significantly more precision than the many existing methods. This work conveyed a 15% improvement in the top

$L/2$ long-range contacts precision on the 150 PSICOV test datasets trained with 3456 proteins from the DeepCov dataset.

TripletRes [15] is another deep learning-based method to predict protein contact maps. The important edge of TripletRes is its capability to learn and directly connect a triplet of co-evolutionary matrices driven from the metagenome and whole-genome databases. It helps to minimize information loss during contact model training. TripletRes also gained the highest precision (71.6%) for the top-$L/5$ long-range contact predictions in the contact prediction category of the CASP13 competition.

The state-of-the-art results in terms of precision range from 57%-76%, 70%-86% and 80%-92% for top L, $\frac{L}{2}$, and $\frac{L}{5}$ residue pairs selected based on predicted probability of having contacts between them [183].

Contact map prediction methods listed above are mainly DNN based. For example, they have used ResNet [13, 14, 119, 121], CNN [107, 108, 117, 121], Dialated CNN [102], Deep Belief Network [118, 176], U-net [177], LSTM-BRNN [14], and ensembles of DNNs [184]. Input features used in these methods include PSSM [14, 102], HMM [14], Covariance Matrix [102, 107, 108], CCM-Pred [14, 102, 116–119], FreeContact [102, 119, 122], ShannonEntropy [12], Contact Potential [12], Precision Matrix [108, 120], Pseudolikelihood Maximization Matrix [120], Compressed Covariance-Matrix [121], and Reduced Precision Matrix [108, 121], and PSICOV [185]. Besides datasets available from the CASP, the other datasets used in evaluating these methods include DNCON2 dataset [119], DeepCov dataset [107], PSICOV150 [185], CAMEO [136], and SPOT-1D [95].

Overall, a broad range of input features such as PSSM, sequence profiles, covariance matrix, precision matrix, secondary structure, and solvent accessibility has been critical for contact map prediction. Likewise, many research groups have employed different Deep Learning frameworks, including ResNet, GANs, FCNs, and U-Nets. Authors observe that all best-performing approaches such as TripletRes [15] and RaptorX [13] use ResNet and its variants among those architectures. The use of deeper MSAs as inputs feature is the key to high performance in many recent approaches for contact prediction. Nevertheless, after the CASP13 competition, most PSP researchers switched from contact prediction to real-valued distance prediction (distogram prediction) as they give much richer information than mere binary contacts for model construction.

12.2 Further Research Challenges

Contact map prediction is largely driven by the belief that the evolution of functional and structural properties of residues in close proximity takes place in sync. To perform such evolutionary coupling analysis (ECA) of residue pairs, MSA is used with respect to a given sequence library. ECA methods include CCMPred [116], FreeContact [122], GREMLIN [186], PlmDCA [187] and PSICOV [185]. Moreover, ECA based meta predictor MetaPSICOV [12]

and NeBCON [188] combines several complementary ECA methods. Nevertheless, ECA methods prove useful in predicting contacts between residue pairs that are far apart in the amino acid sequence. However, ECA methods do not work well in terms of better contact prediction accuracy when fewer homolog proteins are found by MSA [189]. Moreover, the indirect coevolutionary coupling is another problem in which case residues A and C are not in contact but appear to be coevolutionary with each other since they both are separately coevolutionary with respect to B.

Deep learning-based contact map prediction methods, e.g. DeepConPred [176], exploit their ability to learn underlying relationships, particularly when the availability of known proteins is ever increasing. Such other methods, e.g. DeepCov [107] and PconsC4 [177], also use MSA as input features. When MSA based features are not effective because of low availability of homolog proteins, various other features have been used by other contact prediction methods such as RaptorX-Contact [13], DeepContact [117], DeepConPred2 [118], DNCON2 [119], DEEPCON [102], SPOT-Contact [14], DeepCDpred [178], ResPRE [108] and MapPred [121]. Among these methods, RaptorX-Contact [13] and DNCON2 [119] are the first to use complete proteins as the context for prediction and capturing global interaction among residues. Note that earlier contact prediction methods have made predictions for just single residue pairs and have used fixed-sized windows around the residue pairs to capture local interactions. Nevertheless, DNCON2 [119], SPOT-Contact [14], MapPred [121] and TripletRes [15] use multiple types of DNN either in ensembles or in multiple levels. Further, DNN-based methods include DeepECA [190] that uses an end-to-end approach and GANcon [191] that uses GAN.

Overall, contact map prediction accuracy has been dramatically influenced so far by DNNs and ECA [192]. On the one hand, DNNs can capture global coevolutionary coupling patterns from the ever-increasing availability of known proteins. So DNN-based approaches are able to predict more hydrophobic interactions [183]. On the other hand, MSA helps exploit coevolutionary information from many proteins, of which structures are not known. As such ECA based approaches can help predict more salt bridges and disulphide bonds between residues that are far in the amino acid sequence [183].

Contact map prediction has inherent drawbacks [183]. Contact maps cannot distinguish distances that are larger than 8Å. So residues that do not have sufficient numbers of contact residues around might not be properly placed within the 3D structure of a protein. So in such a case, contact map-based PSP methods might not work well. This is one reason for current PSP methods not performing well in the pocket areas or active sites where drug molecules can dock on the proteins.

Another problem with contact maps is that contact maps are not balanced since more than 92% residue pairs are typically not in contact [178]; this unbalanced data potentially creates problems in machine learning. The incorporation of many memory-intensive features in deep learning makes contact prediction costlier. The use of more known and unknown proteins as they

are increasingly available might make contact prediction even costlier. Contact predictors do not perform well when sufficient homologs are not found for MSA-based features. Lastly, new features might be needed for better accuracy since most recent methods rely on similar sets of features. In this direction, instead of using MSA-based features, raw MSA could be used as features.

13 Distance Map Prediction

Distance maps are closely related to contact maps. Both contact and distance maps are translation and rotation invariant two-dimensional projections of 3D protein structures. So many comments on contact maps hold on distance maps. To be specific, *distance maps* are two-dimensional symmetric numeric arrays representing distances between residue pairs. In this case, each residue in the distance computation is typically represented by the C^{β} atom (C^{α} for Glycine). Nevertheless, distances between residues that have long sequence separation can capture long-range global interactions. Such distances are essential for protein structure construction but are challenging to predict, too. Fig. 10 shows the actual residue-residue distance map of the protein with PDB ID 1CF7 predicted by trRosetta [20]. Predicted distance maps could be used as geometric constraints during the construction or refinement of the 3D structures of a protein by using optimisation-based search algorithms [20, 21, 83, 193]. Like predicted contact maps, predicted distance maps could also be used as input features of deep learning-based predictors for other aspects, e.g. in the EMA [194–196].

Distance maps essentially address the data imbalance problem of contact maps in which about 92% residue pairs are not in contact [124]. Moreover, distances are real numbers, while contacts are Boolean values using a certain distance as the threshold. Distance maps can precisely represent inter-residue distances, while contact maps cannot effectively compare residue pairs based on their distances, mainly when both pairs are in contact or are not in contact. Recently distance maps are more informative than contact maps [20, 21, 83, 197]. So, CASP also has introduced a challenge category for distance maps [32].

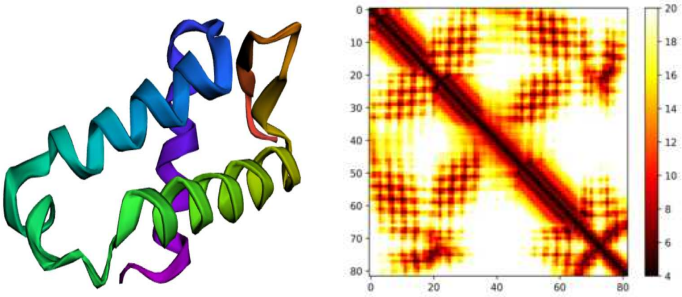

Fig. 10 The three dimensional structure of a protein 1CF7 (left) and its corresponding distance map (right) predicted by trRosetta [20].

13.1 Existing Distance Prediction Methods

Recent distance map prediction methods include RaptorX [83], AlphaFold [21], GANProDist [198], trRosetta [20, 31], PDNET [99], DeepDist [120], AttentiveDist [199], trRosettaX [200], D-QUARK [201], SDP [202] and MDP [203].

Distance map prediction results should ideally be compared using MAE or MSE. However, they are often evaluated using contact map evaluation criteria, and for this contact, probabilities are often computed by using a formula $p = d/4$ if $d > 4.0$ else 1.0 where d is the predicted distance between a residue-pair and p is the probability of having contact between the same residue pair. The underlying assumption is that residue pairs not in contact are not very useful; this assumption might not be justified. Nevertheless, the state-of-the-art distance map prediction results are 1Å-4Å in MAE when only distances up to 16Å are considered. In terms of translated contact map performance, prediction precision range from 45%-50%, 58%-62%, 69%-75% for top L, $\frac{L}{2}$, and $\frac{L}{5}$ residue pairs based on the predicted probability of having contacts between them.

Distance map prediction methods largely follow the contact map prediction methods and so are mostly based on deep learning models. For example, they have used GAN [198], CNN [172], ResNets [20, 120], residual convolutional network [22, 120], and fully convolutional residual network [20, 83, 99, 204]. Input features used in these methods include PSSM [20, 99], contact density [99, 123, 124], predicted contact maps [32], Shannon entropy [99], covariance matrices [99], and precision matrices [99]. The datasets used in evaluating these methods include PSICOV150 dataset [32], DeepCov Dataset [107], CAMEO-HARD dataset [20, 99].

Distance maps were introduced first as a multi-class distance bin prediction problem [123]. Further methods [20, 21, 83, 199] have predicted distances as bins while some other methods [99, 120, 198] have predicted as real values. In the CASP13, the top-ranked models [21, 172, 204, 205] used contact or distance predictions to construct template-free structure to achieve significant progress. After that, the inter-residue distance prediction techniques have become a centre of interest for PSP.

Distance maps have been shown to have helped the PSP method achieve native-like structures [124]. RaptorX [205] extends distance maps by considering distances between C^β atoms of the residue pairs to distances between other atoms of the residue pairs. For example, distances between $C^\alpha - C^\alpha$ atom pairs, $C^\gamma - C^\gamma$ atom pairs (oxygen or sulphur if no C^γ), $C^\alpha - C^\gamma$ atom pairs, and $N - O$ atom pairs. The accuracy levels for all these types of atom pairs need to be improved.

The predicted long or short distances could be translated into protein tertiary structures by web server tools such as trRosetta [20], DMPfold [145] and CONFOLD2 [206]. trRosettaX [200] is an enhanced version of trRosetta as it involves a new multi-scale network, i.e., Res2Net, for the prediction of inter-residue geometries (distance and orientations). trRosettaX also uses

an attention-based module to manipulate multiple homologous templates to enhance accuracy further. Compared with trRosetta, trRosettaX enhances the performance of predicted distances by 6% and 8% on the free modelling category of CASP13 and CASP14, respectively. D-QUARK [201] is an extension of QUARK [28] for deep learning based distance-map and orientation-map predictions. D-QUARK participated as the "QUARK" group in CASP14 and ranked as the top automated server for free modelling targets. SDP [202] predicts inter-residue real distance by scrutinising deep learning models using only two coevolutionary and three non-coevolutionary features. MDP [203] predicts various ranges of real-valued distances by combining distinct deep learning models using a stacked meta-ensemble technique.

Overall, protein distance/contact prediction was mostly driven by two technologies: the residue-residue coevolutionary analysis for informative input feature generation and different deep learning techniques for effectively extracting protein distance/contact patterns from the features [207]. All recent inter-residue distance prediction methods use ResNet as the core deep learning architecture. The RaptorX method [205] wins the title in the CASP13 competition for distance prediction. Since DeepMind's AlphaFold, the more accurate approach, did not contest in the contact prediction classification of the CASP13, it is hard to assess the distance map generated by AlphaFold in terms of contact precision. The trRosetta [20, 31] method developed after the CASP13 competition is outperformed the top CASP13 predictors. Both trRosetta and Alphafold methods use techniques to predict inter-residue orientations, which can help to improve the 3D PSP. Although all of these approaches are publicly accessible, trRosetta is the easiest method to use as it is a totally TensorFlow-based implementation and does not need any input feature generation when MSAs are ready. Most recently, some methods such as DeepDist [120] were developed to predict real-valued inter-residue distances using regression-based deep learning approaches, in addition to multiple distance interval classification. Moreover, the attention mechanism that can choose relevant information in the input features was also applied to predict protein contact maps [208]. In the CASP14 competition, the attention mechanism was also employed by AlphaFold2 [24, 146], tFold [209] and MULTICOM [207] distance predictors to improve distance/structure prediction.

13.2 Further Research Challenges

Despite recent focus and significant progress, distance map prediction still has many challenges. Predicting both long and short distances with similar accuracy levels is challenging. Various types of loss functions have been tried with deep learning methods to put emphasis on long distances, but more effective methods are needed. Also, it is not known whether very long distances (e.g. larger than 16Å) are useful in protein structure construction or refinement. Influenced by contact maps, short distances (smaller than 16Å) are typically used. However, pocket areas where drug molecules could dock on a protein

tend to have larger inter-residue distances. So the prediction of larger distances is actually needed. This somewhat explains why current high-quality PSP methods still struggle in the pocket areas. Further, as is the case with contact prediction, the incorporation of many memory-intensive features in deep learning makes distance prediction costlier. Using more known and unknown proteins as they are increasingly available might make distance prediction even costlier. Also, distance predictors do not perform well when sufficient homologs are not found for MSA-based features. Lastly, new features might be needed for better accuracy since most recent methods rely on similar sets of features. In this direction, instead of using MSA-based features, raw MSA could be used as features.

14 Angle Map Prediction

The angle maps represent various *angle orientations* between residue pairs. Fig. 11 shows five angle orientations $\Omega_{ij} = \Omega_{ji}$, Θ_{ij}, Θ_{ji}, Φ_{ij}, Φ_{ji} between two non-consecutive residues with indexes i and j. Among these angles, Ω and Θ angles are dihedral angles, while Φ angles are planar angles. To be clearer, dihedral angle Ω_{ij} is defined by atoms C_i^α, C_i^β, C_j^β, and C_j^α; dihedral angle Θ_{ij} by atoms N_i, C_i^α, C_i^β, and C_j^β; and dihedral angle Θ_{ji} by atoms N_j, C_j^α, C_j^β, and C_i^β. Nevertheless, for each of the angle types Ω, Θ, and Φ, we can have an *angle map* comprising a two-dimensional non-symmetric array holding the respective angle values for the residue pairs. After contact and distance map predictions and their effectiveness in improving protein structure construction and refinement, inter-residue angle map predictions have emerged to capture further geometric constraints [210, 211]. Many of the comments made on contact or distance map prediction also hold on to angle map prediction. So far predicted angle maps have been used in optimisation-based search for protein structure construction and refinement [20, 199, 212–214]. Predicted angle maps could also be used as input features of deep learning-based predictors for other PSP aspects.

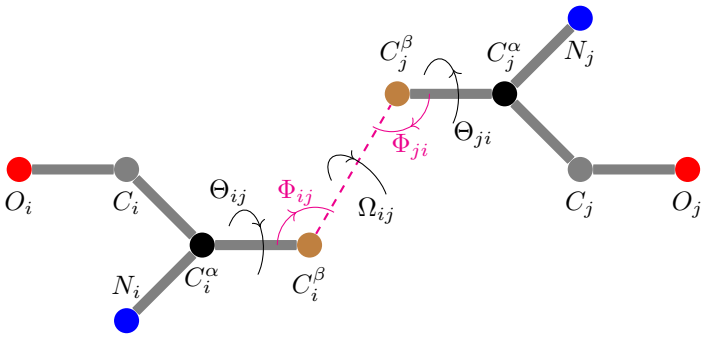

Fig. 11 Angles $\Omega_{ij} = \Omega_{ji}$, Θ_{ij}, Θ_{ji}, Φ_{ij}, Φ_{ji} between two non-consecutive residues with indexes i and j

14.1 Existing Angle Map Prediction Methods

Recent angle map prediction methods include trRosetta [20], trRosettaX [200] and AttentiveDist [199]. These methods have used residual convolutional networks with MSA-based features and predict angle values in bins rather than in actual values. The state-of-the-art prediction accuracy for these methods is not very high, indicating further room for improvement.

14.2 Further Research Challenges

Further challenges for angle map prediction are similar to those for contact and distance maps. It is unknown whether angles between residues far apart in the structures are helpful for protein structure construction or refinement. Input features used in deep learning for angle map prediction are memory intensive, and the use of more proteins as they are increasingly available might make angle prediction even costlier. Further, angle predictors might not work well when sufficient homologs are not found for MSA-based features. Furthermore, new features might be needed for further progress as the existing methods use similar features. In this direction, instead of using MSA-based features, raw MSA could be used as features.

15 Covalent Bond Prediction

Covalent bonds such as disulphide and hydrogen bonds capture long-range interaction among residues. These bonds help stabilise a protein's 3D structure. Below we briefly explore the literature on disulphide and hydrogen bond prediction.

15.1 Disulphide Bond Prediction

Disulphide bonds are covalent bonds between sulphur atoms of cysteine residues. However, only oxydised cysteins, not the reduced ones, form such bonds. Disulphide bonds increase favourable enthalpy interactions in the folded state of a protein and decrease the entropy of the unfolded states and thus stabilise a protein structure [215]. Using predicted disulphide bonds as constraints helps obtain better protein structures [216]. Recent Disulphide bond prediction methods include DBCP [217], Cyscon [218], CRiSP [219], diSBPred [220], and two other methods [221, 222]. These algorithms are mainly machine learning based and have used support vector machine [217], nearest neighbour algorithm [220, 221], support vector regression [218]. The dataset used in these methods include PDBCYS [218, 222], IVD-54 [218, 222], PDB [219], SysPTM [220, 221], and UniProt [220, 221]. Input features used in these methods include PSSM [220–222], Local Similarity Matrix (LSM) [222], Position Specific Estimated Energy (PSEE), residue profile, physiochemical profile, conservation profile, structural profile and flexibility profile [220]. The state-of-the-art results are

about 82% accuracy for predicting whether a cysteine residue forms a disulphide bond and 94% accuracy for predicting a bond between a pair of cysteine residues.

15.2 Hydrogen Bond Prediction

Hydrogen bonds between polar amino and carboxyl groups, in turn, give rise to secondary structures such as helices and sheets. Sheet construction is comparatively harder than helix construction. Usually, two paired sheet residues form hydrogen bonds with each other or with their neighbours [223] and contribute to the stabilisation of the corresponding sheet. Consequently, neural network-based sheet pairing algorithms have been developed, and the use of predicted hydrogen bonds as constraints has improved sheet construction accuracy significantly [215]. Recent hydrogen bond prediction methods include MESHI [224], HAAD [225], and D-I-TASSER [214]. 3Drefine [226, 227] uses MESHI predicted hydrogen bonds and iteratively optimises an atomic-level scoring function. Rosetta software has been used to construct protein structures with significant numbers of sheets by exploiting predicted hydrogen bonds as constraints [228]. D-I-TASSER use DeepPotential [229] to predict hydrogen bonds and incorporates the predicted bonds as constraints in its structural fragment assembly process.

16 Structure Accuracy Prediction

Given a set of protein structures, estimation of their accuracy values with respect to the native structure and then just selecting one is challenging [230, 231]. This problem is also known as EMA and model-quality assessment. EMA research is very recent. The performance of an EMA method depends on both the set of given structures and the metric used to rank them. EMA methods use model-quality features such as stereo-chemical correctness, atomic statistical potential [139, 210, 232], atomic solvent accessibility, secondary structure agreement, and residue-residue contacts [233]. Usually, combinations of multiple model-quality features provide better estimation [233–235]. EMA method QDeep [196] uses distance estimations from DMPfold [145]. PSP refinement methods GalaxyRefine2 [236], RefineD [237], and Baker suite [195] incorporates EMA into their refinement pipeline. Other PSP methods DMPfold2 [125] and AlphaFold2 [24, 146] uses EMA as an integral part of their sequence-to-model pipeline.

17 Abstract Local Structure Prediction

In proteins, structures are more conserved than sequences [80]. Once secondary structures are formed, the amino acid sequences within the secondary structures are perhaps not very important for the global interactions and the overall structure of the entire protein. Super secondary structures and secondary structural motifs are abstract local structures comprising secondary structures

in various arrangements, e.g. α-helix hairpins, β-hairpins, coiled coils, Greek key motif, Rossmann fold, α-turn-α, α-loop-α, β-α-β [238]. Identifying such abstract structures and reasoning about them would allow one to focus on the overall shape of a protein's structure before delving into the finer details, particularly for large and complex proteins.

17.1 Existing Abstract Local Structure Prediction Methods

Machine learning algorithms have been developed for super secondary structure prediction [239–245]. Machine learning techniques used by these methods include SVM [239], Logistic Regression (LogReg), Extra Trees, Random Decision Forest (RDF), K Nearest Neighbor (KNN), and Gradient Boosting Classifier (GBC) [244, 245]. Many machine learning based methods just predict only one type super secondary structures e.g. coiled coils by SpiriCoil [246], RFCoil [240], AAFreqCoil [242], CCBuilder2.0 [243] while other methods predict multiple types e.g. [239, 241]. StackSSSPred [244] is a stacking-based machine learning method for the prediction of two types of β-hairpins and β-α-β structures. A recent method [245] predicts β-β motifs, β-α motifs, α-β motifs, and α-α motifs using the random forest technique and obtains 72%-84% accuracy for ArchDB dataset.

17.2 Further Research Challenges

To identify various abstract local structures, a key step would be to find the relative positions of the secondary structures. This would allow one to reason about protein structures more at the secondary structure level than at the residue level. Note that protein secondary structure prediction mainly assigns the secondary structure type to each residue in a protein. The issue here is to find the position of one secondary structure with respect to another. This new PSP sub-problem could be named as *protein secondary structure position prediction* problem. This would essentially allow working with super secondary structures and secondary structural motifs. This would also help develop scalable PSP approaches.

18 Energy Function Design

Since energy functions (also known as physics-based energy functions) are not precisely known for protein structures, designing such functions remains a great challenge. One key requirement of such energy functions is that they must be differentiable since computational approaches will minimise them for a given protein. Amino acids and their constituent atoms essentially have fundamental chemical and physical characteristics, including electrostatics, quantum mechanics, and statistical mechanics. So the energy functions for proteins will essentially have various physical and chemical components. For example, Van der Waals forces, electrostatic forces, bond energies are used in various

force field functions that include CHARMM [42], AMBER [43], GROMOS [247], OPLS-AA [248], and Rosetta [249]. These methods have been under development for a long time and have made significant progress. In general, these methods are computationally very intensive since they are at the atomic level and involve interactions among atoms. Moreover, energy functions are recomputed for any change made in the structures during the search. Machine learning algorithms, however, do not explicitly depend on energy functions. Given that machine learning algorithms for PSP have obtained significant success recently, energy function-based PSP would gradually come into focus since the basic hypothesis that "the sequence alone determines the structure" has not yet been properly addressed. We think in light of the recent progress, some simplified forms of energy functions could be explored.

19 Refinement Search using Energy Functions

Such energy-based refinement search methods do not use scoring functions that are designed using information extracted from other proteins. For example, such methods do not use information such as APL [81] or fragment assembly [250] or predicted distance or contact or angle maps. These methods rather depend on the physicochemical properties of the amino acids and force field-based energy components as are used in CHARMM [42] or AMBER [43]. For high-resolution structure prediction, many other characteristics that include side chains, hydrogen bonds, and hydrophobic burial are to be used in energy functions [249, 251].

19.1 Existing Energy-Based Refinement Search Methods

Recent energy-based refinement search algorithms are mainly population-based. A bi-objective adaptive differential evolution algorithm named ADE-MO/D [252] uses bonded and non-bonded components of the CHARMM [42] energy function. 3Drefine method [227] uses atomic-level energy minimisation using a combination of physics and knowledge-based force fields, and ReFOLD [253] uses molecular dynamics-based refinement search. An evolutionary algorithm MO3 [254] uses the bonded and non-bonded components of CHARMM [42] but also adds SASA as another energy component. A multi-objective self-adaptive differential evolution algorithm named MODCSA-CA [255] uses the bonded and non-bonded components of CHARMM [42] but includes side-chain angles as well. Another algorithm named AIMOES [256] is like MO3 in the search method and the energy function but includes hydrogen atoms in its all-atom models. A particle swarm optimisation algorithm named MOPSO [257] uses the CHARMM and the dFIRE [258] energy functions. A memetic algorithm [259] uses the Rosetta [249] energy function. A recent method [260] employs different niching methods into the memetic algorithm to obtain better RMSD.

19.2 Further Research Challenges

Overall, energy-based refinement search approaches still face significant challenges in proteins having a length larger than 150 residues [36]. Two key reasons are the compute-intensive energy functions and the search methods that mainly adopt randomised neighbourhood generation approaches. However, sacrificing atomic details and considering only fundamental atoms will require further simplified structural models. Note that there is a body of literature on lattice-based simplified models and related energy or scoring functions, but considering the focus of the paper, we only very briefly discuss them in Section 24. On the other hand, it would be interesting if *ab initio* search algorithms could replace their randomised approaches with more focused approaches in neighbourhood generation. Perhaps, the inferior parts of the current conformations could be identified, and the neighbouring conformations could be generated, making changes in those identified parts.

20 Scoring Function Design

Scoring functions are proxy energy functions (known as knowledge-based energy functions). They are preferably continuous-valued, computationally cheaper, and yet informative enough in terms of distinguishing the native structure from other decoy structures of a protein. Like energy functions, scoring functions should be differentiable since they will perhaps be optimised by continuous-space search algorithms.

20.1 Existing Scoring Functions

Early scoring functions include hydrophobic polar (HP) [261] and Miyazawa-Jernigan (MJ) matrix [262, 263] functions. HP is based on the hydrophobicity property of each amino acid. Hydrophobic residues normally hide at the core of a protein, while the polar residues stay at the surface. MJ matrix is based on the statistical observations of the interactions among 20 types of residues. The HP model, the MJ matrix, and their variants have been mainly used in lattice-based PSP.

Further statistics based scoring functions include dDFIRE [258], RWplus [210], GOAP [211], and OPUS-DOSP [138]. Among these, dDFIRE ("dipolar" DFIRE) [258] is a classic energy function based on dipole-dipole orientation angles. Note that dDFIRE considers each polar atom as a dipole, with the orientation of the dipole defined by the bond vectors that bind the polar atom to other heavy atoms. Based on the distance between two atoms and the three angles involved in dipole-dipole interactions, the dDFIRE energy function is determined from protein structures. This method considers the hydrogen bonding interaction via the physical dipole-dipole interaction. More crucially, it gives a consistent treatment for putative orientation-dependent interactions between polar and non-polar atoms, as well as between non-hydrogen-bonded

polar atoms. Furthermore, an integrated approach of distance and angle dependence yields a statistical energy function that is parameter-free. RW is a pair-wise distance-dependent, atomic statistical potential function that uses a random-walk chain as a reference state, which was optimised by incorporating a new side-chain orientation-dependent energy term to form RWplus [210]. GOAP [211] is a generalized orientation-dependent all-atomic statistical potential function for PSP. OPUS-DOSP [138] is a distance and orientation-based all-atomic potential generated from side-chain packing.

Recent scoring functions mainly comprise machine learning predicted geometric constraints. For example, predicted contact maps [264–266], predicted distance maps [20, 21, 267] [22, 193], predicted angle maps [20, 22, 267], and predicted hydrogen bonds [264] have been used. Some scoring functions combine the predicted constraints with components from energy functions such as CHARMM [42] and Rosetta [249]. Many scoring functions have multiple components, and such scoring functions normally take a weighted sum of the components [268]. Finding a suitable weighting scheme is a significant AI challenge.

20.2 Further Research Challenges

From the HP model to the MJ matrix model to contact or distance or angle maps, we see an interesting evolution of scoring functions based on interactions among generic two types of residues to generic twenty types of residues to protein and position-specific residues. When predicted contacts or distances, or angle orientations are used, one key concern in scoring function design is to penalise a given conformation for the deviations from the predicted values. Various penalty functions have been used so far for predicted contacts. Such penalty functions include modified Lorentz potential [269], soft square [206], square well [172, 270], bounded potential [172], and CGLFOLD [265]. No such penalty function is found for distance map-based scoring functions since real distances are mostly converted to contacts by existing PSP refinement methods. However, for predicted binned distances, bin probabilities have been captured by differentiable functions using spline transformation [20, 21], and then those functions are minimised. When multiple components have been used in a scoring function, a general trend is to take the sum of the components. However, relative scaling of the components is an important issue in this case, and if a weighted sum is used, determining the weights becomes difficult. Considering each component as a separate objective, Pareto optimality or interleaving techniques could be used.

21 Refinement Search using Scoring Functions

Given a protein's amino acid sequence, scoring-based search approaches extract information from other proteins or their structures, presumably using machine learning algorithms and exploiting the information in the conformational search for the given protein. The extracted information could be in terms of

predicted main or side chain angles, predicted secondary structures and their associated angle ranges, APLs, fragments collected from other proteins, and predicted contact or distance or angle maps.

21.1 Existing Scoring-Based Search Algorithms

Recent hybrid search algorithms are mainly population-based metaheuristic algorithms. A genetic algorithm named GA-APL [81] uses genetic algorithms with the Rosetta [249] energy function and generates neighbours using APLs. Another genetic algorithm [268] uses 3DIGARS and SASA as energy functions. A multi-stage memetic algorithm combines the Rosetta method as a local search standard to examine the low-energy conformation [112]. Another memetic algorithm [148] hybridises with a simulated annealing algorithm using ad-hoc crossover and mutation operators for PSP with the Rosetta energy function and APL [81] in neighbour generations. Two other memetic algorithms [150, 271] use Rosetta and SASA as energy functions and APLs in a neighbour generation. CONFOLD [264] converts inter-residue contacts and secondary structures into the distance, dihedral angle, and hydrogen bond constraints and uses the Crystallography and NMR system (CNS) [272]. CONFOLD2 [206] is like CONFOLD, but uses a soft square energy function. MODE-K [267] is a multi-objective differential evolution algorithm that uses distance and angle orientation based energy function RWplus [210]. CGLFold [265] uses fragment assembly based global exploration, loop perturbation based differential evolution algorithm for local exploration, and contact maps and Rosetta3 [273] as energy functions. SPOT-Fold [266] uses predicted contact maps, predicted dihedral angles, dDFIRE [258], soft-square energy functions, and the CNS system.

Very recent scoring-based search algorithms that use distance and angle maps obtain outstanding results. These algorithms are behind the recent success of PSP. AlphaFold [21] predicts distance maps as distance bins and converts them into differentiable energy potentials, and minimises their sum with a gradient descent algorithm. Then, trRosetta [20] uses predicted inter-residue angle and distance maps and the Rosetta energy function within the Rosetta-based gradient descent optimisation framework. A combined search method [274] uses predicted distances from AlphaFold with physics-based refinement via molecular dynamics simulations. MULTICOM2 [193] uses predicted distance maps with the DFOLD [120] system, which is like the CNS [275] system. RaptorX-3DModeling [22] uses distances between various atoms from residue pairs and angle orientations between residue pairs and then converts the predicted bins into differential energy potentials and performs gradient descent minimisation.

21.2 Further Research Challenges

Overall, hybrid search approaches exploit the best potentials of machine learning and search-based optimisation algorithms. It will be interesting to see

what other characteristics could be learned from known proteins and then exploited in the refinement search. However, the search algorithms, regardless of the population-based or optimisation algorithms used, mainly adopt randomised neighbourhood generation approaches. So replacing the randomised approaches with more focused approaches in neighbourhood generation would be a meaningful direction from AI perspective. In this context, the inferior parts of the current conformations could be identified, and the neighbouring conformations could be generated, making changes in those identified parts.

22 Fragment Assembly Approaches

Fragment assembly approaches extract short (3–15 residues) and contiguous fragments from known protein structures and build a fragment library [28, 250]. Then, during protein structure construction or refinement of a protein structure, for a given local window of residues, the fragment library is searched to find the fragments with an identical amino acid sequence or identical secondary structures. The structural properties (e.g. dihedral angles) of the matching fragments are then assigned to the local window, one fragment at a time, and the fragment resulting in the best scoring value is accepted. The selection of a local window, in this case, is guided by an iterative optimisation search algorithm, e.g. Monte Carlo simulations or population-based or local search algorithms.

22.1 Existing Fragment Assembly Approaches

The performance of a fragment assembly approach depends on the quality of the fragment library. Recent fragment library construction approaches are Rosetta [276], QUARK [28], FRAGFOLD [250], NNMAKE [277], Flip [278], Profrager [279], FRAGSION [280], LRFragLib [281], DeepFragLib [282], and CGLFold [265]. Rosetta [276] is one of the best classic fragment assembly-based methods that uses simple simulated annealing Monte Carlo simulation approach to assemble native-like structures from fragments of unrelated protein structures with similar local sequences utilising Bayesian scoring functions. Rosetta scores fragments of three and nine residues based on profile-profile and secondary structure similarity between the query sequence and fragments over a specified window size. Subsequent versions of Rosetta have shown significant improvement in the performance [283, 284]. QUARK [28] is another classic TFM structural prediction method based on a similar principle of fragment assembly employing Monte Carlo simulation versions but with various techniques for fragment generation and energy function design. QUARK is built on continuous fragment assembly using both knowledge and physics-based energy terms. A number of novel energy terms and Monte Carlo movements are used to enhance the efficiency of both force field and search engine. Rosetta and QUARK have been consistently ranked among the top free-modelling approaches in CASP competitions since CASP3 and CASP9, respectively.

FRAGFOLD employs super-secondary structure fragments to convey the correlation between consecutive secondary structure elements. NNMAKE is a Rosetta-based approach for fragment library generation. NNMake uses a fragment scoring approach. Profrager uses PSIPRED to give a confidence score. Flip makes use of the predominant predicted secondary structure of fragments to improve the precision of driven fragment libraries. FRAGSION uses Hidden Markov Model to select fragments. LRFragLib uses logistic regression models, while DeepFragLib uses deep learning models. Among recent fragment assembly approaches, CGLFold [265] uses fragment assembly in its global exploration phase along with the contact maps and the Rosetta [249, 276] energy function. Another method proposed in [285] uses a gradient minimisation algorithm with predicted inter-residue distances and backbone angles. C-QUARK [286] is a recent extension of QUARK [28] and incorporates several deep-learning and coevolution-based contact-maps to guide the replica-exchange Monte Carlo fragment assembly simulations. The tested results of C-QUARK on 64 free-modelling targets from the CASP13 data demonstrate that C-QUARK achieves an average GDT score that was 5% higher than the best CASP predictors.

22.2 Further Research Challenges

Note that APL [81] could be viewed as single-residue fragments. Also, fragment assembly approaches are essentially hybrid search algorithms with neighbourhood generation using fragments. In this case, the search space is restricted by the fragment library. As observed in hybrid search algorithms and fragment assembly approaches, the randomised selection of local windows is a key issue besides fragment library construction. Instead of using a random selection approach, any informed approach to selecting the local windows will be an exciting direction from AI perspective.

23 Side Chain Prediction

Protein side-chain prediction deals with the side chains of the amino acids, assuming the main chain angles have already been predicted, and so the protein's main chain has already been constructed. However, both main and side chains could also be predicted in a combined fashion, and this would arguably be more challenging than the two-step process. As noted before, protein side chains can be represented by using χ angles. Side-chain angle prediction research is so far dominated by search-based algorithms [287]. So more deep learning approaches for χ angle prediction could be investigated in future.

23.1 Existing Side Chain Prediction Approaches

Recent side-chain prediction methods include FASPR [288] and OPUS-Rota3 [289]. Both methods deal with side-chain packing assuming main chain folding has already been done. Another recent method does folding and packing

in an interleaving fashion [290] using the FastDesign protocol of Rosetta. The state-of-the-art results in side-chain prediction are about 50% in terms of accuracy with a tolerance limit of 20Å, and about 71%-87% accuracy for various dihedrals angles with a tolerance limit of 40Å.

Search-based side-chain angle prediction involves a rotamer library, a scoring function, and a search strategy. A *rotamer library* contains side-chains derived from high-resolution X-ray structures. Some rotamer libraries contain backbone-independent side chain information, while some other libraries contain backbone-dependent information. Arguably backbone-dependent rotamer libraries lead to better accuracy. A *scoring function* for side-chain angle prediction is typically molecular dynamics based. It contains components involving electrostatics, hydrogen bonding, solvation free energy, van der Waals forces, and least simple steric energy [291–293]. Molecular dynamics-based scoring functions are quite sensitive to slight changes in atom-atom distances and inevitably lead to much higher repulsive energies relative to the native structures. Knowledge-based scoring functions do not exhibit such characteristics. The Knowledge-based scoring function could include rotamer frequency, side-chain orientation, number of neighbouring contact points, and contact surface area. Knowledge-based scoring functions have been designed using neural networks. For example, SIDEpro [294] uses a shallow neural network to learn inter-atomic distances, and OPUS-RotaNN [289] uses a DNN to predict side-chain candidates from the rotamer library. Nevertheless, clash detection is an important scoring criterion in side-chain packing [291, 292]. A *search algorithm* for side-chain angle prediction trades off between speed and accuracy. Moreover, deterministic algorithms are often preferred over stochastic ones. Examples of deterministic search algorithms are integer linear programming [295], mixed-integer linear programming [296], branch-and-bound [297], graph-theoretic algorithms [298], and tree decomposition [288, 299]. Examples of nondeterministic algorithms include Monte Carlo [139], simulated annealing [300], and genetic algorithms [301]. Early search algorithms such as CIS-RR [291] and RASP [292] sample continuous search space, while recent search algorithms such as OPUS-Rota2 [302] and FASPR [288] uses sample space discretised in terms of angular values or the number of samples available in the rotamer library.

23.2 Further Research Challenges

Search-based algorithms used in side-chain packing mainly adopt the randomised selection of the side chains to undergo changes. From AI perspective, as we have already discussed for main chain prediction, it is important to make informed selection decisions. So it would be an interesting direction if side chains that are really in bad shape could be selected by using heuristics.

24 Simplified Protein Structure Prediction

The focus of this review paper is not on simplified PSP. While deep learning methods have been heavily used in recent days in detailed PSP, metaheuristic algorithms to date have been heavily used in simplified PSP. So considering the coverage of the metaheuristic algorithms, we briefly discuss simplified PSP.

24.1 Simplified Protein Representations

Simplified representations of protein structures mainly deal with C^{α} atoms of the amino acid residues and hence only with the main chains. Basically, the C^{α} atoms are considered to be representing the amino acid residues. Also, one key assumption is two C^{α} atoms of two successive residues are at a constant Euclidean distance. Simplified PSP models could be lattice-based or off-lattice. The lattice-based models put angular restrictions on the placement of successive amino acid residues with respect to the preceding amino acid residues, while the off-lattice models do not. So lattice-based conformations could be represented by integer coordinates while off-lattice models need real-valued coordinates. Although the actual PSP is to find 3D protein structures, some simplified PSP models, both lattice-based and off-lattice, are defined over the two-dimensional space. Moreover, some simplified PSP models, e.g. square and cubic lattices, allow only 90° angles between every three successive residues. The 3D lattices used in PSP include cubic, hexagonal closed packs, and face-centred cubic (FCC). However, considering popularity, we discuss only the FCC lattice model. We also discuss one 3D off-lattice simplified PSP model.

3D FCC Lattice

In comparison to other 3D lattices such as the cubic or the body-centred cubic, the 3D FCC lattice has the highest average density [303]. The amino acids are situated in the centre and middle of the edges of the cubic unit cell in the FCC lattice. As a result, each lattice point has 12 neighbours with 12 basis vectors. A sequence of n amino acids in a protein can be encoded as a sequence of $n-1$ basis vectors that define the 3D shape. In the FCC lattice, two points $p=(x,y,z)$ and $q=(x',y',z')$ are adjacent in the lattice if and only if $|x-x'| \leq 1$, $|y-y'| \leq 1$, $|z-z'| \leq 1$ and $|x-x'|+|y-y'|+|z-z'|=2$.

3D Off-Lattice Model

The 3D off-lattice model [304] is the same as the $\theta-\tau$ based representation discussed in Section 7 and Fig. 8. It needs $n-2$ bond angles between every three successive residues and $n-3$ dihedral angles among every four successive residues.

24.2 Simplified Scoring Functions

Below, we discuss the HP model and the MJ matrix scoring functions, although we have briefly discussed them in Section 20 with other scoring functions. Below, we also discuss the AB model and Berrera Matrix.

HP and AB Models

The HP model uses the hydrophobicity property to classify 20 amino acids into two classes: hydrophobic and hydrophilic. When two non-consecutive hydrophobic amino acid residues come in contact, the pair release one unit of energy. The objective is to minimise the total free energy released by all possible non-consecutive hydrophobic amino acid residue pairs that are in contact. The HP model is one of the simplest scoring functions and could be used to find the hydrophobic core of a protein. There are a number of variants of the HP model. The HP model has been typically used with the lattice-based protein structure representation models. With off-lattice representation models, an HP-like scoring model is called *AB model*.

MJ and Berrera Matrices

The Miyazawa-Jernigan (MJ) Matrix [263] is a 20×20 matrix denoting interaction potential between each possible pair of 20 amino acids. The interaction potentials are obtained by performing a statistical analysis of known protein structures. This model is based on the assumption that the average characteristics of residue-residue contacts created in a large number of protein crystal structures immediately correspond to actual differences in interactions among residues, as if the amino acid sequence of each protein, as well as intra-residue and short-range interactions, had no significant contribution [305]. In the MJ matrix, a residue is represented by its side-chain centre, which for Glycine is taken as the position of its C^{α} atom. Moreover, a pair of residues is defined to be in contact if they are not nearest neighbours in the sequence and the distance between their centres is less than 6.5Å. The protein structures can be used to count the number of various types of residue-residue contacts [306]. *Berrera Model* [307] is another 20×20 matrix obtained by calculating empirical contact energies between 20 types of amino acid residues but using a different set of proteins and quasi-chemical approximation.

24.3 Simplified PSP Approaches

Existing simplified PSP approaches ranges over search algorithms that include tabu search, genetic algorithms, memetic algorithms, simulated annealing algorithms, and particle swarm optimisation algorithms. The search algorithms listed above are mainly perturbative metaheuristic algorithms. As such, they suffer from a re-visitation of the same conformations again, getting stuck at local minima and stagnating around plateaus. With a given scoring function, the main challenge in simplified PSP models is to explore the astronomically large search space, even when the lattice models use discretisation.

Lattice-Based PSP Methods:

A random-walk-based stagnation recovery method [308] tries to get rid of the local minima problem in PSP using HP model and FCC lattice. A hybrid search approach [309] combines a tabu local search technique within a genetic algorithm for the HP model and FCC lattice. A genetic algorithm [310] with FCC lattice employs a high-resolution MJ matrix model for PSP and a low-resolution HP energy model in focusing the search towards exploring structures that have hydrophobic cores. A genetic algorithm named GAPlus [311] uses a hydrophobic core-directed macro-mutation operator to intensify the search and a duplication elimination strategy to avoid early convergence on FCC lattice with HP Model. An ant colony optimisation algorithm [312] for PSP uses FCC lattice with HP model and MJ matrix. A systematic evolutionary optimisation technique called multimodal memetic framework [151] explores FCC lattice space using the HP model. A statistical method [313] uses Bayesian learning for cubic lattice and HP model. A quantum genetic algorithm [314] finds optimised solutions for PSP on a square lattice and HP model.

24.4 AB Off-Lattice Model:

An improved simulated annealing algorithm [315] for simplified PSP. A recent simplified PSP method [316] combines deep learning and numerical optimisation to analyse the optimal morphology of protein structures. A multi-peak sampling method [317] studies the stability of protein structures. Another simplified PSP method [318] uses tabu search in generating initial solutions and particle swarm optimisation in the construction of neighbourhood functions.

24.5 Discussion

Simplified models are toy models considering both protein structure representations and scoring functions. A study [319, 320] shows that simplified models using cubic or FCC lattice exhibit negative correlations between RMSD and the scoring functions based on HP model or MJ or Berrera matrices. Moreover, lattice-based models result in coarse-grain protein structures. While research with these models has some academic value, they rather carry very little significance considering the actual PSP problem with 3D dihedral angle or cartesian coordinate-based formulations and with the recent development of machine-learning based protein specific scoring functions. However, the off-lattice $\theta - \tau$ based models, with their ability to better represent main chains, carry much significance, although not with simplified scoring functions but rather with recent protein scoring functions.

25 Recent Outstanding CASP Methods

PSP research has been largely driven by the double-blind CASP [78] competition. CASP has been taking place every two years since 1994. CASP has many tracks to assess the performance of various computational methods for

PSP. CASP2 separated the TBM and TFM tracks. We refer to the CASP website (https://predictioncenter.org) for its excellent resource repository. By CASP11, the accuracy levels that the best performing TFM methods could achieve were at a plateau. However, in CASP13, a jump in accuracy was seen due to the addition of deep-learning approaches for contact prediction. In the CASP13, for the first time, the DeepMind's AlphaFold modelled exceptionally challenging targets with an average GDT_TS of 70% [21].

In CASP14, a further leap in structure prediction accuracy was seen with the best model for any target level reaching a GDT_TS above 90%, a range at the level of laboratory experimental accuracy [90]. As expected, in CASP14, AlphaFold2 produced the most accurate models for most protein targets, and AlphaFold2 models were capable of solving the vast majority of crystal structures. However, the second-best approach, an improved version of the deep learning-based trRosetta [20], outperformed the best model AlphaFold [21] in CASP13, indicating the role of community-based bench-marking experiments in driving the further improvement of the field. Table 5 shows the top 10 3D protein structure predictors in CASP14.

A recent method named RoseTTAFold [23] uses a deep-learning algorithm with features inspired by AlphaFold2 at CASP14. Many ideas of AlphaFold2 were independently reproduced and implemented in RoseTTAFold. In addition, ColabFold [33], a free PSP platform coupled with Google Colaboratory, offers an accelerated prediction of protein structures and complexes by combining the fast homology search of Many-against-Many sequence searching (MMseqs2) [321] with AlphaFold2 and RoseTTAFold. Table 6 provides a description of recent outstanding methods with available links to access the resources.

Table 5 Top 10 predictors in CASP14 PSP competition.

Group Name	Sum Z-score (> 0.0)	Avg TM-score	Avg GDT_TS
AlphaFold2	244.0217	0.9052	0.8801
BAKER	92.1241	0.7388	0.6695
BAKER-experimental	91.4731	0.7334	0.6653
FEIG-R2	74.5627	0.7088	0.6464
Zhang	68.8922	0.7142	0.6386
tFold_human	65.2157	0.7021	0.6280
MULTICOM	64.0531	0.6989	0.6302
QUARK	62.9711	0.6959	0.6234
Zhang-Server	62.9122	0.6978	0.6249
tFold-IDT_human	62.0795	0.6862	0.6179

25.1 AlphaFold and AlphaFold2

AlphaFold [21] and AlphaFold2 [24] achieved remarkable success in CASP13 and CASP14 with AlphaFold2 achieving near 90% GDT scores.

AlphaFold [21] presented a deep-learning approach to PSP. The central component of AlphaFold is a CNN. It demonstrated that a learned, protein-specific potential could be constructed by training deep neural networks to

Table 6 Recent outstanding methods in PSP with available links to access the resources.

Methods	Resources
AlphaFold [21]	https://github.com/deepmind/
RoseTTAFold [23]	https://github.com/RosettaCommons/RoseTTAFold
trRosetta [20]	https://yanglab.nankai.edu.cn/trRosetta/ Webserver - http://bioinf.cs.ucl.ac.uk/psipred/
D-QUARK [201]	https://zhanglab.dcmb.med.umich.edu/D-QUARK/
RaptorX [22]	http://raptorx.uchicago.edu
DMPfold [125]	https://github.com/psipred/DMPfold Webserver - http://bioinf.cs.ucl.ac.uk/psipred/
D-I-TASSER [214]	https://zhanglab.dcmb.med.umich.edu/D-I-TASSER/
TripletRes [15]	https://zhanglab.ccmb.med.umich.edu/TripletRes/
ColabFold [33]	https://github.com/sokrypton/ColabFold

produce accurate protein structure prediction given its sequence and predicting the structure by reducing the potential using gradient descent. Backbone torsion angles and inter-residue distances are among the predictions made by the neural network. Distance predictions offer more specific structural information than contact predictions and give a richer training input for the proposed neural network. The network may propagate distance information that takes into account covariation, local structure, and the identities of neighbouring residues by jointly predicting various distances. Combining the predicted probability distributions yields a simple, principled protein-specific potential.

AlphaFold2 [24] predicts the structure of individual protein chains by using information from the amino acid sequence, MSA, and homologous structures. The neural network's core component, known as Evoformer, comprises a neural representation of the MSAs and pairwise relationships between distinct amino acids in the protein. A set of other neural network modules combines and processes these two forms of neural representations. The pairwise representation of the relation contains information about the relative positions of amino acids in the given chain. This format is then used to predict the relative distances between the amino acids in the chain using a binned distance distribution (distogram). To predict the final structure, the MSA embedding is then combined with the pair embedding. The neural network model is trained end-to-end with gradients propagating from the predicted structure through the entire network. By using the information in MSAs of related proteins as the raw input features for end-to-end training, AlphaFold2 was capable of predicting the 3D atomic coordinates of folded protein structures at a median GDT test score of 92.4% in the most recent round of the CASP Assessment (CASP14).

26 Overall Discussion and Further Challenges

PSP is a holy grail in many fields that include computational biology, computational chemistry, and bioinformatics. PSP is of paramount importance for many reasons, but its use in drug design is greatly worth mentioning. In vivo and in vitro methods for PSP are significantly costlier, very slow, and error-prone. PSP is computationally challenging, even with simplified structural representations and energy functions. The hardness is further compounded

by more realistic continuous-valued angular representations and atomic-level energy functions. There are many unknown challenges in PSP, but computational PSP has still achieved significant progress over the last 60 years. More categorically, great progress has been made in recent years through the use of deep learning techniques in predicting protein-specific dihedral angles, contact maps, distance maps, angle orientations, and hydrogen bonds and then using them within another layer of end-to-end learning via attention-based network models. Below we discuss the implications of the recent success and the challenges that lie ahead. However, we do that from AI perspective.

26.1 AI Progress Behind Recent PSP Success

To drive the PSP research, the biennial CASP [78] competition has played a pivotal role for the last three decades. In these three decades, computing technology has also gone through tremendous progress. Very high-performance computers are now available with greater memory and speed. Powerful graphical processing units in modern computers have increased computational capabilities even further. With this hardware progress, the AI fields, in general, also have obtained a notable boost. Moreover, machine learning, particularly deep learning, has gone through a revolution. Complex deep learning models could now capture complicated input-output functional relations. This was not previously possible because of the unavailability of suitable hardware. Complex deep learning models need massive data as well. In PSP, the steady growth in the numbers of proteins with known and unknown structures has made a massive volume of data available. With all these, the progress in global connectivity via the internet has paved the way for easy and effective cross-fertilisation of researchers from all over the world, both in terms of ideas and methods. Lastly, big research industries have also invested in the PSP area to take the research to a different level. The result is apparent in the outcome of the recent CASP competition, where AlphaFold2 [24] by Google's DeepMind has achieved outstanding success.

26.2 Success of AlphaFold and PSP Community

We observe that AlphaFold greatly enjoys recent success in obtaining high-quality protein structures. Its end-to-end training of attention networks that can selectively extract rich information from sequences in MSA and its iterative structural refinement based on local structure error estimation has made a great difference. Given that each research effort takes the state-of-the-art a step ahead of other related workarounds, the question is, what is the extent of AlphaFold's leap from the scientific point of view? The PSP community has developed many effective techniques over the decades. Existing techniques that are an essential part of AlphaFold include end-to-end learning, distance predictions, attention networks, the use of raw MSA, and iterative improvements. Given that AlphaFold has not taken PSP research in a completely different direction, we are of the opinion of attributing the scientific success to a large

extent with the whole PSP community. Of course, from an engineering point of view, AlphaFold, with its industry-level expertise, has demonstrated a fine tapestry in combining many effective techniques together. Also, AlphaFold has exploited its extensive computing resources that are most likely beyond the reach of most academic researchers. Nevertheless, the actual impact of the results achieved by AlphaFold will critically depend on the complete availability of its program and source code in the public domain. Otherwise, the community needs to continue its scientific journey from where it still stands at.

26.3 Recent Progress from AI Perspectives

From AI perspective, a problem is considered solvable when it can be solved theoretically regardless of the computational resources used. AI interests go beyond just determining whether a problem is solvable and extend to seeking efficient and alternative solutions that could produce high-quality solutions using practical computational resources, particularly for computationally intractable problems. In PSP, we observe that PSP research was making slow but gradual progress and also, at the same time, perhaps was having some level of doubt whether very accurate protein structures could cause at all be predicted, particularly when the underlying energy functions are not known precisely. In this respect, AlphaFold, of course, has provided a renewed supply of enthusiasm and optimism that the PSP problem is indeed solvable up to an incredible accuracy level. Further, from a bioinformatics perspective, PSP is a problem in which a computational solution is needed to achieve very high prediction accuracy levels, but there is no significant concern in the type of computational method being used. This is, of course, the general view of any research area that appears to be an application for AI. However, from AI perspective, the method itself, along with its type, is interesting. Given that the recent success in PSP is largely driven by the neuralisation of many key steps in the PSP pipeline, a renewed AI interest would be to seek search-based approaches. The advantages of search-based approaches over machine learning or neural network-based approaches are, of course, a better understanding of how the methods internally work, in other words, the explainability of the method.

26.4 Data Driven vs Knowledge Driven PSP Approaches

"More data leads to better models" is more or less known in AI. Also, more or less known in AI is that "more computation leads to better solutions". The "bitter lesson" blog by Computer Scientist Richard Sutton mentions that so far in history, "general methods that leverage computation are ultimately the most effective, and by a large margin", and he also mentions the reason is Moore's law, meaning hardware acceleration achieved over the years. One inherent reason behind such facts could be that the way human solves problems is not necessarily the way computers solve problems. Nevertheless, the AI community overall is divided over data vs knowledge in the definition of objective functions

and exhaustiveness vs randomness vs informedness in decision making. The data-driven AI has relied on the exploitation of voluminous noisy data and has taken probabilistic approaches in decision making. The knowledge-driven AI has relied on the well-defined crispy representation of knowledge, perhaps captured from data or coded by experts and has taken deterministic approaches in decision making. The exhaustive approaches provide a theoretical guarantee of obtaining optimal solutions but prove practically futile. The random approaches, while often working well, are debated whether they at all represent intelligence. So informed approaches are often preferred since they provide explainability and better understanding to humans. We observe that the recent PSP progress is mostly coming from data-driven approaches in machine learning and probabilistic and random approaches in search and optimisation. So from AI perspective, reducing the data dependency and increasingly relying on knowledge would be an interesting direction in PSP. One can argue at this point that although the deep learning approaches use huge data, they eventually capture knowledge extracted from the data. While this is somewhat correct, we, however, consider deep learning models only as hazy intermediate representations that are beyond human understanding. What we mean by knowledge is that knowledge would be a crispy well-defined analytical formula that clearly makes sense to a human. In terms of the general model mentioned by Richard Sutton, we see that current PSP methods are still largely designed by humans and are not general since human intuitions are largely coded in the program. We can say that because the type of techniques (e.g. distance maps, end-to-end approach, etc.) are basically coming from human experts, and a machine is not automatically determining the type of models. In our view, a general method would take the least amount of input from humans in its model construction.

26.5 Where AI Can Contribute More

In this review paper, we mainly discuss single domain PSP. Despite AlphaFold2's achievement, there still exists room for improvement in PSP. AlphaFold2 obtains experimental accuracy for only about 35% template-free proteins. For other proteins, a very high accuracy (< 0.5Å) level is yet to be achieved. Moreover, drug designing experts are particularly concerned that the predicted protein structures at the current accuracy levels are not yet reliable in the active and allosteric pocket areas where drug molecules could dock [45]. So prediction accuracy needs to be improved in the pocket areas. The computational resources used by AlphaFold2 are also beyond the reach of many research groups. Minimising the resource requirement is a crucial challenge. Simplifying AlphaFold2's entire method or developing alternative simpler methods without sacrificing accuracy results is another challenge. Moreover, "the sequence determines the structure" still largely remains unaddressed as the current success mainly depends on the exploitation of knowledge learnt from known and unknown proteins in the form of MSA-based features and machine learning

approaches in general. Below we list a number of subproblems where further research effort is needed.

1. Long Distance Prediction: Current distance prediction methods mainly learn and predict short inter-residue distances. However, pocket areas of the proteins normally have longer distances, and current PSP methods do not perform well in those areas.
2. Learning New Constraints: Current PSP methods use predicted contacts and distance maps, and angle orientations. Finding any other type of geometric constraint could be very useful.
3. Intelligence Sample Generation: Current PSP search methods generate conformations by making changes in randomly selected parts of current conformations. More constraint-guided sampling is needed, particularly in the parts that have low accuracy.
4. Ablation of Input Features: Current machine learning approaches for PSP tend to use large numbers of input features. So models are to be developed that use fewer features and preferably fewer MSA-based coevolutionary features.
5. Reduction of Training Data: Current machine learning approaches for PSP use huge numbers of known and unknown training proteins. Reducing the number of training proteins, both known and unknown, would be useful.
6. Using Simpler Learning Models: Current deep learning methods for PSP tend to use large ensembles of complex neural networks. Reducing the architectural complexity of learning models would be a good direction.
7. Improving Search Methods: Combining multiple scoring functions is challenging, particularly when scoring functions are subject to error and only partially capture an unknown ideal energy function. Better many-objective search algorithms could be useful.
8. Search Methods using Basic Hypothesis: Given that machine learning-based approaches have almost solved PSP, improving search methods that do not or minimally exploit information from other proteins would be really interesting.
9. Split and Merging in Deep Learning: Deep learning methods often need a fixed-sized input. However, proteins have variable numbers of residues. So padding is needed for proteins smaller than the model's size. Unfortunately, padding introduces noise. A potential way is to split a protein into many pieces, perhaps overlapping pieces, and use them in deep learning. However, the challenge is then to merge them and obtain resultant predictions for entire proteins reconciling predictions made for pieces.
10. Iterative Use of Predicted Proteins: Iterative computational methods improve from one iteration to another, exploiting the result obtained in the previous iteration. Given then highly accurate predicted structures are now available, it would be interesting to investigate whether further improvement could

be achieved by using predicted structures as input for any prediction model since they are native-like structures.

11. Secondary structure prediction by search: Physics-based energy functions exist for entire proteins but not for secondary structures. If such energy functions could be developed for secondary structures, search methods for PSP would be greatly advanced.

26.6 Further AI Approaches for PSP

1. Reinforcement Learning: Reinforcement learning is a machine learning training method. Reinforcement learning rewards desired behaviours and punish undesired ones. Reinforcement learning is based on trial and error and is useful when no objective function is known beforehand. Given that PSP has no clearly defined energy functions developing reinforcement learning approaches for PSP would be an exciting research direction. This will essentially be a more generalised model.
2. Explanation Based Learning: Explanation-based learning exploits a formal definition of a problem in explaining the success and failure of a training example and generalising the explanation knowledge. Even with Adhoc problem models, as long as they are formally defined, explanation-based learning can help perform supervised learning with fewer training examples. So developing explanation-based learning for PSP would be another exciting research direction from AI perspective.
3. Abstraction Based Approaches: PSP approaches have already taken an abstraction-based approach when residues have been represented by C^{α} atoms only. This has reduced computational complexity greatly, particularly in the size of the search space and in the computation of the scoring functions. Abstraction-based approaches could be extended to secondary structures, super secondary structures and super secondary structural motifs. These approaches need adequate research attention.

26.7 What AI can Learn from PSP

One great challenge PSP poses to AI is in the energy function. As noted before, most AI methods rely on the availability of a precisely defined objective function. When such an objective function is costlier, sometimes surrogate functions are used, and some other times auxiliary objective functions are used. However, what should be done in a scenario where the objective function is not known at all is a key AI challenge. So AI researchers can learn a lot from PSP in dealing with hazily defined objective functions and perhaps when multiple such hazy definitions are possible, and also multiple of them are to be used. Given that each hazy function will have associated errors, another key AI challenge is to reconcile them since they could be more accurate in some cases while the others could be less accurate. These are somewhat different from Pareto optimisation.

27 Conclusion

In this review paper, we provide an overview of the current state-of-the-art of template-free protein structure prediction (PSP) research. We survey the literature methodologically and comprehensively using two bibliometric analysis tools, Gephi and Bibexcel. Template-free PSP has recently obtained significant success via deep learning and search-based optimisation methods. However, to obtain protein structures that could be effectively used in drug discovery, more sophisticated and advanced artificial intelligence (AI) techniques are needed. Unfortunately, AI researchers struggle to get into PSP research because of the lack of a comprehensive computational view of PSP along with the related research challenges. Moreover, existing PSP review papers cover PSP research at a very high level and only some parts of PSP and only from a particular singular viewpoint. We fill in the knowledge gap by covering background and literature on relevant sub-problems of template-free PSP. From our long experience in AI and PSP, we analyse PSP from AI perspectives and point out potential research directions.

28 Competing interests

There is NO Competing Interest.

29 Author contributions statement

M.M.M.M. and M.A.H.N. contributed equally in all parts of the work and are joint-first authors. A.S. took part in discussions and reviewed the manuscript.

30 Acknowledgments

This work is partly supported by the Australian Research Council Discovery Grant DP180102727 and the AHEAD OPERATIONS Project of Sri Lanka.

References

[1] Jana, N.D., Das, S., Sil, J.: A Metaheuristic Approach to Protein Structure Prediction. Springer, Gewerbestrasse 11, 6330 Cham, Switzerland (2018)

[2] Cavanagh, J., Fairbrother, W.J., Palmer III, A.G., Skelton, N.J.: Protein NMR Spectroscopy: Principles and Practice. Academic press, United States (1996)

[3] Glusker, J.: X-ray crystallography of proteins. Methods of Biochemical Analysis, 1–72 (2009)

[4] Comellas, G., Rienstra, C.M.: Protein structure determination by magic-angle spinning solid-state NMR, and insights into the formation, structure, and stability of amyloid fibrils. Annual review of biophysics **42**, 515–536 (2013)

[5] Bagaria, A., Jaravine, V., Güntert, P.: Estimating structure quality trends in the Protein Data Bank by equivalent resolution. Computational biology and chemistry **46**, 8–15 (2013)

[6] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank. Nucleic acids research **28**(1), 235–242 (2000)

[7] Bairoch, A., Bougueleret, L., Altairac, S., Amendolia, V., Auchincloss, A., Puy, G.A., Axelsen, K., Baratin, D., Blatter, M.-C., Boeckmann, B., *et al.*: The universal protein resource (uniprot). Nucleic acids research **36**, 190–195 (2008)

[8] Dotu, I., Cebrian, M., Van Hentenryck, P., Clote, P.: On lattice protein structure prediction revisited. IEEE/ACM Transactions on Computational Biology and Bioinformatics **8**(6), 1620–1632 (2011)

[9] Dal Palu, A., Dovier, A., Fogolari, F., Pontelli, E.: Exploring protein fragment assembly using CLP. In: IJCAI, pp. 2590–2595 (2011)

[10] Anfinsen, C.B.: Principles that govern the folding of protein chains. Science **181**(4096), 223–230 (1973)

[11] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A., Wheeler, D.L.: Genbank. Nucleic acids research **28**(1), 15–18 (2000)

[12] Jones, D.T., Singh, T., Kosciolek, T., Tetchner, S.: Metapsicov: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics **31**(7), 999–1006 (2015)

[13] Wang, S., Sun, S., Li, Z., Zhang, R., Xu, J.: Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology **13**(1), 1005324 (2017)

[14] Hanson, J., Paliwal, K., Litfin, T., Yang, Y., Zhou, Y.: Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics **34**(23), 4039–4045 (2018)

[15] Li, Y., Zhang, C., Bell, E.W., Zheng, W., Zhou, X., Yu, D.-J., Zhang, Y.:

Deducing high-accuracy protein contact-maps from a triplet of coevolutionary matrices through deep residual convolutional networks. PLoS computational biology **17**(3), 1008865 (2021)

[16] Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., De Groot, B.L., Grubmüller, H., MacKerell, A.D.: Charmm36m: an improved force field for folded and intrinsically disordered proteins. Nature methods **14**(1), 71–73 (2017)

[17] Park, H., Bradley, P., Greisen Jr, P., Liu, Y., Mulligan, V.K., Kim, D.E., Baker, D., DiMaio, F.: Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. Journal of chemical theory and computation **12**(12), 6201–6212 (2016)

[18] Heo, L., Feig, M.: Experimental accuracy in protein structure refinement via molecular dynamics simulations. Proceedings of the National Academy of Sciences **115**(52), 13276–13281 (2018)

[19] Park, H., Ovchinnikov, S., Kim, D.E., DiMaio, F., Baker, D.: Protein homology model refinement by large-scale energy optimization. Proceedings of the National Academy of Sciences **115**(12), 3054–3059 (2018)

[20] Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., Baker, D.: Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences **117**(3), 1496–1503 (2020)

[21] Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W., Bridgland, A., *et al.*: Improved protein structure prediction using potentials from deep learning. Nature **577**(7792), 706–710 (2020)

[22] Xu, J., Mcpartlon, M., Li, J.: Improved protein structure prediction by deep learning irrespective of co-evolution information. Nature Machine Intelligence, 1–9 (2021)

[23] Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al.: Accurate prediction of protein structures and interactions using a three-track neural network. Science (2021)

[24] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al.: Highly accurate protein structure prediction with alphafold. Nature, 1–11 (2021)

[25] Kelley, L.A., Sternberg, M.J.: Protein structure prediction on the web: a case study using the phyre server. Nature protocols **4**(3), 363–371 (2009)

[26] Roy, A., Kucukural, A., Zhang, Y.: I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols **5**(4), 725–738 (2010)

[27] Wang, Z., Eickholt, J., Cheng, J.: Multicom: a multi-level combination approach to protein structure prediction and its assessments in casp8. Bioinformatics **26**(7), 882–888 (2010)

[28] Xu, D., Zhang, Y.: 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)

[29] Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., Xu, J.: Template-based protein structure modeling using the raptorx web server. Nature protocols **7**(8), 1511–1522 (2012)

[30] Zheng, W., Li, Y., Zhang, C., Pearce, R., Mortuza, S., Zhang, Y.: Deep-learning contact-map guided protein structure prediction in CASP13. Proteins: Structure, Function, and Bioinformatics **87**(12), 1149–1164 (2019)

[31] Du, Z., Su, H., Wang, W., Ye, L., Wei, H., Peng, Z., Anishchenko, I., Baker, D., Yang, J.: The trrosetta server for fast and accurate protein structure prediction. Nature protocols **16**(12), 5634–5651 (2021)

[32] Adhikari, B., Shrestha, B., Bernardini, M., Hou, J., Lea, J.: DISTEVAL: a web server for evaluating predicted protein distances. BMC bioinformatics **22**(1), 1–9 (2021)

[33] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., Steinegger, M.: ColabFold: making protein folding accessible to all. Nature Methods, 1–4 (2022)

[34] Høie, M.H., Kiehl, E.N., Petersen, B., Nielsen, M., Winther, O., Nielsen, H., Hallgren, J., Marcatili, P.: NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. Nucleic Acids Research (2022)

[35] Jayaram, B., Bhushan, K., Shenoy, S.R., Narang, P., Bose, S., Agrawal, P., Sahu, D., Pandey, V.: Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins. Nucleic acids research **34**(21), 6195–6204 (2006)

[36] Dhingra, S., Sowdhamini, R., Cadet, F., Offmann, B.: A glance into the

evolution of template-free protein structure prediction methodologies. Biochimie (2020)

[37] Khor, B.Y., Tye, G.J., Lim, T.S., Choong, Y.S.: General overview on structure prediction of twilight-zone proteins. Theoretical Biology and Medical Modelling **12**(1), 1–11 (2015)

[38] Liwo, A., Khalili, M., Scheraga, H.A.: Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proceedings of the National Academy of Sciences **102**(7), 2362–2367 (2005)

[39] Klepeis, J.L., Wei, Y., Hecht, M.H., Floudas, C.A.: Ab initio prediction of the three-dimensional structure of a de novo designed protein: A double-blind case study. Proteins: Structure, Function, and Bioinformatics **58**(3), 560–570 (2005)

[40] Bradley, P., Misura, K.M., Baker, D.: Toward high-resolution de novo structure prediction for small proteins. Science **309**(5742), 1868–1871 (2005)

[41] Jauch, R., Yeo, H.C., Kolatkar, P.R., Clarke, N.D.: Assessment of casp7 structure predictions for template free targets. Proteins: Structure, Function, and Bioinformatics **69**(S8), 57–67 (2007)

[42] Brooks, B.R., Brooks III, C.L., Mackerell Jr, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., *et al.*: CHARMM: the biomolecular simulation program. Journal of computational chemistry **30**(10), 1545–1614 (2009)

[43] Case, D.A., Cheatham III, T.E., Darden, T., Gohlke, H., Luo, R., Merz Jr, K.M., Onufriev, A., Simmerling, C., Wang, B., Woods, R.J.: The amber biomolecular simulation programs. Journal of computational chemistry **26**(16), 1668–1688 (2005)

[44] Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., *et al.*: Highly accurate protein structure prediction for the human proteome. Nature **596**(7873), 590–596 (2021)

[45] Mullard, A.: What does alphafold mean for drug discovery? Nature reviews. Drug discovery (2021)

[46] Mignan, A., Broccardo, M.: One neuron versus deep learning in aftershock prediction. Nature **574**(7776), 1–3 (2019)

[47] AlQuraishi, M.: Machine learning in protein structure prediction. Current Opinion in Chemical Biology **65**, 1–8 (2021)

[48] Jisna, V., Jayaraj, P.: Protein structure prediction: Conventional and deep learning perspectives. The Protein Journal, 1–23 (2021)

[49] Wardah, W., Khan, M.G., Sharma, A., Rashid, M.A.: Protein secondary structure prediction using neural networks and deep learning: A review. Computational biology and chemistry **81**, 1–8 (2019)

[50] Torrisi, M., Pollastri, G., Le, Q.: Deep learning methods in protein structure prediction. Computational and Structural Biotechnology Journal (2020)

[51] Bastian, M., Heymann, S., Jacomy, M.: Gephi: an open source software for exploring and manipulating networks. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 3(1) (2009)

[52] Persson, O., Danell, R., Schneider, J.W.: How to use bibexcel for various types of bibliometric analysis. Celebrating scholarly communication studies: A Festschrift for Olle Persson at his 60th Birthday **5**, 9–24 (2009)

[53] Kuhlman, B., Bradley, P.: Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology **20**(11), 681–697 (2019)

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

[55] Pearce, R., Zhang, Y.: Deep learning techniques have significantly impacted protein structure prediction and protein design. Current Opinion in Structural Biology **68**, 194–207 (2021)

[56] Pearce, R., Zhang, Y.: Toward the solution of the protein structure prediction problem. Journal of Biological Chemistry, 100870 (2021)

[57] Biehn, S.E., Lindert, S.: Protein structure prediction with mass spectrometry data. Annual review of physical chemistry **73**, 1–19 (2022)

[58] Lee, D., Xiong, D., Wierbowski, S., Li, L., Liang, S., Yu, H.: Deep learning methods for 3D structural proteome and interactome modeling. Current Opinion in Structural Biology **73**, 102329 (2022)

[59] Mittal, A., Jayaram, B., Shenoy, S., Bawa, T.S.: A stoichiometry driven universal spatial organization of backbones of folded proteins: are there chargaff's rules for protein folding? Journal of Biomolecular Structure

and Dynamics **28**(2), 133–142 (2010)

[60] Jiang, Q., Jin, X., Lee, S.-J., Yao, S.: Protein secondary structure prediction: A survey of the state of the art. Journal of Molecular Graphics and Modelling **76**, 379–402 (2017)

[61] Kabsch, W., Sander, C.: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules **22**(12), 2577–2637 (1983)

[62] Blum, C., Roli, A.: Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM computing surveys (CSUR) **35**(3), 268–308 (2003)

[63] Bianchi, L., Dorigo, M., Gambardella, L.M., Gutjahr, W.J.: A survey on metaheuristics for stochastic combinatorial optimization. Natural Computing **8**(2), 239–287 (2009)

[64] Shonkwiler, R.W., Mendivil, F.: Explorations in Monte Carlo Methods. Springer, Switzerland (2009)

[65] Goldberg, D.E.: Genetic algorithms in search. Optimization, and MachineLearning (1989)

[66] Glover, F.W., Kochenberger, G.A.: Handbook of Metaheuristics vol. 57. Springer, Switzerland (2006)

[67] Talbi, E.-G.: Metaheuristics: from Design to Implementation vol. 74. John Wiley & Sons, Hoboken, New Jersey, U.S (2009)

[68] Hussain, K., Salleh, M.N.M., Cheng, S., Shi, Y.: Metaheuristic research: a comprehensive survey. Artificial Intelligence Review **52**(4), 2191–2233 (2019)

[69] Khanduja, N., Bhushan, B.: Recent advances and application of metaheuristic algorithms: A survey. Metaheuristic and Evolutionary Computation, 207 (2020)

[70] Mufassirin, M.M., Ragel, R.G.: A novel filter-wrapper based feature selection approach for cancer data classification. In: 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), pp. 1–6 (2018). IEEE

[71] Fielding, A.H.: An introduction to machine learning methods. In: Machine Learning Methods for Ecological Applications, pp. 1–35. Springer, Switzerland (1999)

[72] Smialowski, P., Martin-Galiano, A.J., Cox, J., Frishman, D.: Predicting

experimental properties of proteins from sequence by machine learning techniques. Current Protein and Peptide Science **8**(2), 121–133 (2007)

[73] Marsland, S.: Machine Learning: an Algorithmic Perspective. CRC press, Boca Raton, Florida (2015)

[74] Somvanshi, M., Chavan, P., Tambade, S., Shinde, S.: A review of machine learning techniques using decision tree and support vector machine. In: 2016 International Conference on Computing Communication Control and Automation (ICCUBEA), pp. 1–7 (2016). IEEE

[75] Vieira, A., Ribeiro, B.: Introduction to Deep Learning Business Applications for Developers. Springer, Switzerland (2018)

[76] Shrestha, A., Mahmood, A.: Review of deep learning algorithms and architectures. IEEE Access **7**, 53040–53065 (2019)

[77] Kugunavar, S., Prabhakar, C.: Convolutional neural networks for the diagnosis and prognosis of the coronavirus disease pandemic. Visual Computing for Industry, Biomedicine, and Art **4**(1), 1–14 (2021)

[78] Moult, J., Pedersen, J.T., Judson, R., Fidelis, K.: A large-scale experiment to assess protein structure prediction methods. Wiley Online Library (1995)

[79] Lavor, C., Alves, R., Figueiredo, W., Petraglia, A., Maculan, N.: Clifford algebra and the discretizable molecular distance geometry problem. Advances in Applied Clifford Algebras **25**(4), 925–942 (2015)

[80] Yang, Y., Gao, J., Wang, J., Heffernan, R., Hanson, J., Paliwal, K., Zhou, Y.: Sixty-five years of the long march in protein secondary structure prediction: the final stretch? Briefings in bioinformatics **19**(3), 482–494 (2018)

[81] Borguesan, B., e Silva, M.B., Grisci, B., Inostroza-Ponta, M., Dorn, M.: APL: an angle probability list to improve knowledge-based metaheuristics for the three-dimensional protein structure prediction. Computational biology and chemistry **59**, 142–157 (2015)

[82] Zhou, X.-g., Zhang, G.-j., Hao, X.-h., Yu, L.: A novel differential evolution algorithm using local abstract convex underestimate strategy for global optimization. Computers & Operations Research **75**, 132–149 (2016)

[83] Xu, J.: Distance-based protein folding powered by deep learning. Proceedings of the National Academy of Sciences **116**(34), 16856–16865 (2019)

[84] Levinthal, C.: Are there pathways for protein folding? Journal de chimie physique **65**, 44–45 (1968)

[85] Zhang, Y., Skolnick, J.: Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics **57**(4), 702–710 (2004)

[86] Zemla, A., Venclovas, Č., Moult, J., Fidelis, K.: Processing and evaluation of predictions in CASP4. Wiley Online Library (2001)

[87] Mariani, V., Biasini, M., Barbato, A., Schwede, T.: lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics **29**(21), 2722–2728 (2013)

[88] Kabsch, W.: A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography **32**(5), 922–923 (1976)

[89] Zemla, A.: LGA: a method for finding 3D similarities in protein structures. Nucleic acids research **31**(13), 3370–3374 (2003)

[90] Pereira, J., Simpkin, A.J., Hartmann, M.D., Rigden, D.J., Keegan, R.M., Lupas, A.N.: High-accuracy protein structure prediction in CASP14. Proteins: Structure, Function, and Bioinformatics **89**(12), 1687–1699 (2021)

[91] Meiler, J., Müller, M., Zeidler, A., Schmäschke, F.: Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. Molecular modeling annual **7**(9), 360–369 (2001)

[92] Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y., Yang, Y.: Predicting backbone $c\alpha$ angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. Journal of computational chemistry **35**(28), 2040–2046 (2014)

[93] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Yang, Y., Zhou, Y.: Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. Scientific reports **5**(1), 1–11 (2015)

[94] Heffernan, R., Yang, Y., Paliwal, K., Zhou, Y.: Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics **33**(18), 2842–2849 (2017)

[95] Hanson, J., Paliwal, K., Litfin, T., Yang, Y., Zhou, Y.: Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. Bioinformatics **35**(14), 2403–2410 (2018)

[96] Fang, C.: Applications of deep neural networks to protein structure prediction. PhD thesis, University of Missouri-Columbia (2018)

[97] Richmond, T.J.: Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect. Journal of molecular biology **178**(1), 63–89 (1984)

[98] Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., Zehfus, M.H.: Hydrophobicity of amino acid residues in globular proteins. Science **229**(4716), 834–838 (1985)

[99] Adhikari, B.: A fully open-source framework for deep learning protein real-valued distances. Scientific reports **10**(1), 1–10 (2020)

[100] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research **25**(17), 3389–3402 (1997)

[101] Görmez, Y., Sabzekar, M., Aydin, Z.: IGPRED: combination of convolutional neural and graph convolutional networks for protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics (2021)

[102] Adhikari, B.: DEEPCON: protein contact prediction using dilated convolutional neural networks with dropout. Bioinformatics **36**(2), 470–477 (2020)

[103] Kinjo, A.R., Nakamura, H.: Nature of protein family signatures: insights from singular value analysis of position-specific scoring matrices. PloS one **3**(4), 1963 (2008)

[104] Remmert, M., Biegert, A., Hauser, A., Söding, J.: HHblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nature methods **9**(2), 173 (2012)

[105] Klausen, M.S., Jespersen, M.C., Nielsen, H., Jensen, K.K., Jurtz, V.I., Soenderby, C.K., Sommer, M.O.A., Winther, O., Nielsen, M., Petersen, B., *et al.*: NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. Proteins: Structure, Function, and

Bioinformatics **87**(6), 520–527 (2019)

[106] Xia, X.: Hidden markov models and protein secondary structure prediction. In: Bioinformatics and the Cell, pp. 145–172. Springer, Switzerland (2018)

[107] Jones, D.T., Kandathil, S.M.: High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics **34**(19), 3308–3315 (2018)

[108] Li, Y., Hu, J., Zhang, C., Yu, D.-J., Zhang, Y.: ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics **35**(22), 4647–4655 (2019)

[109] Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., Weigt, M.: Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences **108**(49), 1293–1301 (2011)

[110] Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., Sander, C.: Protein 3D structure computed from evolutionary sequence variation. PloS one **6**(12), 28766 (2011)

[111] Yanofsky, C., Horn, V., Thorpe, D.: Protein structure relationships revealed by mutational analysis. Science **146**(3651), 1593–1594 (1964)

[112] Garza-Fabre, M., Kandathil, S.M., Handl, J., Knowles, J., Lovell, S.C.: Generating, maintaining, and exploiting diversity in a memetic algorithm for protein structure prediction. Evolutionary computation **24**(4), 577–607 (2016)

[113] Belda, I., Madurga, S., Tarragó, T., Llorà, X., Giralt, E.: Evolutionary computation and multimodal search: A good combination to tackle molecular diversity in the field of peptide design. Molecular diversity **11**(1), 7–21 (2007)

[114] Moscato, P., *et al.*: On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms. Caltech concurrent computation program, C3P Report **826**, 1989 (1989)

[115] Moscato, P., Cotta, C.: A modern introduction to memetic algorithms. In: Handbook of Metaheuristics, pp. 141–183. Springer, Switzerland (2010)

[116] Seemayer, S., Gruber, M., Söding, J.: CCMpred–fast and precise prediction of protein residue–residue contacts from correlated mutations.

Bioinformatics **30**(21), 3128–3130 (2014)

[117] Liu, Y., Palmedo, P., Ye, Q., Berger, B., Peng, J.: Enhancing evolutionary couplings with deep convolutional neural networks. Cell systems **6**(1), 65–74 (2018)

[118] Ding, W., Mao, W., Shao, D., Zhang, W., Gong, H.: DeepConPred2: an improved method for the prediction of protein residue contacts. Computational and structural biotechnology journal **16**, 503–510 (2018)

[119] Adhikari, B., Hou, J., Cheng, J.: DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics **34**(9), 1466–1472 (2018)

[120] Wu, T., Guo, Z., Hou, J., Cheng, J.: DeepDist: real-value inter-residue distance prediction with deep residual convolutional network. BMC bioinformatics **22**(1), 1–17 (2021)

[121] Wu, Q., Peng, Z., Anishchenko, I., Cong, Q., Baker, D., Yang, J.: Protein contact prediction using metagenome sequence data and residual neural networks. Bioinformatics **36**(1), 41–48 (2020)

[122] Kaján, L., Hopf, T.A., Kalaš, M., Marks, D.S., Rost, B.: FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC bioinformatics **15**(1), 1–6 (2014)

[123] Walsh, I., Baù, D., Martin, A.J., Mooney, C., Vullo, A., Pollastri, G.: Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC structural biology **9**(1), 1–20 (2009)

[124] Kukic, P., Mirabello, C., Tradigo, G., Walsh, I., Veltri, P., Pollastri, G.: Toward an accurate prediction of inter-residue distances in proteins using 2D recursive neural networks. BMC bioinformatics **15**(1), 1–15 (2014)

[125] Mirabello, C., Wallner, B.: RAWMSA: end-to-end deep learning using raw multiple sequence alignments. PloS one **14**(8), 0220182 (2019)

[126] Mataeimoghadam, F., Newton, M.H., Dehzangi, A., Karim, A., Jayaram, B., Ranganathan, S., Sattar, A.: Enhancing protein backbone angle prediction by using simpler models of deep neural networks. Scientific Reports **10**(1), 1–12 (2020)

[127] Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., Zhou, Y.: SPIDER2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In: Prediction of Protein Secondary Structure, pp.

55–63. Springer, Switzerland (2017)

[128] Gao, Y., Wang, S., Deng, M., Xu, J.: RaptorX-Angle: real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning. BMC bioinformatics **19**(4), 73–84 (2018)

[129] Fang, C., Shang, Y., Xu, D.: Prediction of protein backbone torsion angles using deep residual inception neural networks. IEEE/ACM transactions on computational biology and bioinformatics **16**(3), 1020–1028 (2018)

[130] Zhong, W., Gu, F.: Predicting local protein 3D structures using clustering deep recurrent neural network. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2020)

[131] Xu, G., Wang, Q., Ma, J.: OPUS-TASS: a protein backbone torsion angles and secondary structure predictor based on ensemble neural networks. Bioinformatics (Oxford, England) (2020)

[132] Heffernan, R., Paliwal, K., Lyons, J., Singh, J., Yang, Y., Zhou, Y.: Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. Journal of computational chemistry **39**(26), 2210–2216 (2018)

[133] Singh, J., Litfin, T., Paliwal, K., Singh, J., Hanumanthappa, A.K., Zhou, Y.: SPOT-1D-Single: improving the single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility and half-sphere exposures using a large training set and ensembled deep learning. Bioinformatics (2021)

[134] Kotowski, K., Smolarczyk, T., Roterman-Konieczna, I., Stapor, K.: ProteinUnet-an efficient alternative to SPIDER3-single for sequence-based prediction of protein secondary structures. Journal of computational chemistry **42**(1), 50–59 (2021)

[135] Wang, G., Dunbrack, R.L.: PISCES: recent improvements to a pdb sequence culling server. Nucleic acids research **33**(suppl_2), 94–98 (2005)

[136] Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., Schwede, T.: The protein model portal—a comprehensive resource for protein structure and model information. Database **2013** (2013)

[137] Wang, G., Dunbrack Jr, R.L.: Pisces: a protein sequence culling server. Bioinformatics **19**(12), 1589–1591 (2003)

[138] Xu, G., Ma, T., Zang, T., Sun, W., Wang, Q., Ma, J.: OPUS-DOSP: a

distance-and orientation-dependent all-atom potential derived from side-chain packing. Journal of molecular biology **429**(20), 3113–3120 (2017)

[139] Lu, M., Dousis, A.D., Ma, J.: OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. Journal of molecular biology **376**(1), 288–301 (2008)

[140] Newton, M., Mataeimoghadam, F., Zaman, R., Sattar, A.: Secondary structure specific simpler prediction models for protein backbone angles. BMC bioinformatics **23**(1), 1–14 (2022)

[141] Görmez, Y., Aydin, Z.: IGPRED-MultiTask: a deep learning model to predict protein secondary structure, torsion angles and solvent accessibility. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2022)

[142] Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., Söding, J.: HH-suite3 for fast remote homology detection and deep protein annotation. BMC bioinformatics **20**(1), 1–15 (2019)

[143] Gao, J., Yang, Y., Zhou, Y.: Grid-based prediction of torsion angle probabilities of protein backbone and its application to discrimination of protein intrinsic disorder regions and selection of model structures. BMC bioinformatics **19**(1), 1–8 (2018)

[144] Ingraham, J., Riesselman, A., Sander, C., Marks, D.: Learning protein structure with a differentiable simulator. In: International Conference on Learning Representations (2018)

[145] Greener, J.G., Kandathil, S.M., Jones, D.T.: Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nature communications **10**(1), 1–13 (2019)

[146] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Žídek, A., Bridgland, A., *et al.*: High accuracy protein structure prediction using deep learning. Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book) **22**, 24 (2020)

[147] Ma, Y., Liu, Y., Cheng, J.: Protein secondary structure prediction based on data partition and semi-random subspace method. Scientific reports **8**(1), 1–10 (2018)

[148] Correa, L., Borguesan, B., Farfán, C., Inostroza-Ponta, M., Dorn, M.: A memetic algorithm for 3D protein structure prediction problem. IEEE/ACM transactions on computational biology and bioinformatics **15**(3), 690–704 (2016)

[149] Narloch, P.H., Parpinelli, R.S.: The protein structure prediction problem approached by a cascade differential evolution algorithm using ROSETTA. In: 2017 Brazilian Conference on Intelligent Systems (BRACIS), pp. 294–299 (2017). IEEE

[150] de Lima Corrêa, L., Dorn, M.: A multi-population memetic algorithm for the 3D protein structure prediction problem. Swarm and Evolutionary Computation **55**, 100677 (2020)

[151] Nazmul, R., Chetty, M., Chowdhury, A.R.: Multimodal memetic framework for low-resolution protein structure prediction. Swarm and Evolutionary Computation **52**, 100608 (2020)

[152] Zaman, R., Newton, M.H., Mataeimoghadam, F., Sattar, A.: Constraint guided neighbour generation for protein structure prediction. IEEE Access (2022)

[153] Newton, M.H., Zaman, R., Mataeimoghadam, F., Rahman, J., Sattar, A.: Constraint guided beta-sheet refinement for protein structure prediction. Computational Biology and Chemistry, 107773 (2022)

[154] Mirabello, C., Pollastri, G.: Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. Bioinformatics **29**(16), 2056–2058 (2013)

[155] Magnan, C.N., Baldi, P.: SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics **30**(18), 2592–2597 (2014)

[156] Wang, S., Peng, J., Ma, J., Xu, J.: Protein secondary structure prediction using deep convolutional neural fields. Scientific reports **6**, 18962 (2016)

[157] Wang, Y., Mao, H., Yi, Z.: Protein secondary structure prediction by using deep learning method. Knowledge-Based Systems **118**, 115–123 (2017)

[158] Fang, C., Shang, Y., Xu, D.: MUFOLD-SS: new deep inception-inside-inception networks for protein secondary structure prediction. Proteins: Structure, Function, and Bioinformatics **86**(5), 592–598 (2018)

[159] Torrisi, M., Kaleel, M., Pollastri, G.: Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. Scientific reports **9**(1), 1–12 (2019)

[160] Dehghani, T., Naghibzadeh, M., Eghdami, M.: BetaDL: a protein beta-sheet predictor utilizing a deep learning model and independent set

solution. Computers in Biology and Medicine **104**, 241–249 (2019)

[161] Rost, B.: Protein secondary structure prediction continues to rise. Journal of structural biology **134**(2-3), 204–218 (2001)

[162] Singh, J., Paliwal, K., Litfin, T., Singh, J., Zhou, Y.: Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment. Scientific reports **12**(1), 1–9 (2022)

[163] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences **118**(15) (2021)

[164] Wang, S., Li, W., Liu, S., Xu, J.: RaptorX-Property: a web server for protein structure property prediction. Nucleic acids research **44**(W1), 430–435 (2016)

[165] Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D., Darnell, J.: Hierarchical structure of proteins. In: Molecular Cell Biology. 4th Edition. WH Freeman, Macmillan Higher Education, US (2000)

[166] Liljas, A., Liljas, L., Lindblom, G., Nissen, P., Kjeldgaard, M., Ash, M.-r.: Textbook of Structural Biology vol. 8. World Scientific, Singapore (2016)

[167] DasGupta, D., Kaushik, R., Jayaram, B.: From ramachandran maps to tertiary structures of proteins. The Journal of Physical Chemistry B **119**(34), 11136–11145 (2015)

[168] Jing, X., Dong, Q., Lu, R., Dong, Q.: Protein inter-residue contacts prediction: methods, performances and applications. Current Bioinformatics **14**(3), 178–189 (2019)

[169] Newton, M.H., Rahman, J., Zaman, R., Sattar, A.: Enhancing protein contact map prediction accuracy via ensembles of inter-residue distance predictors. Computational Biology and Chemistry, 107700 (2022)

[170] Adhikari, B., Cheng, J.: Protein residue contacts and prediction methods. In: Data Mining Techniques for the Life Sciences, pp. 463–476. Springer, Switzerland (2016)

[171] Zhang, G.-J., Ma, L.-F., Wang, X.-Q., Zhou, X.-G.: Secondary structure and contact guided differential evolution for protein structure prediction. IEEE/ACM transactions on computational biology and bioinformatics **17**(3), 1068–1081 (2018)

[172] Hou, J., Wu, T., Cao, R., Cheng, J.: Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. Proteins: Structure, Function, and Bioinformatics **87**(12), 1165–1178 (2019)

[173] Sathyapriya, R., Duarte, J.M., Stehr, H., Filippis, I., Lappe, M.: Defining an essence of structure determining residue contacts in proteins. PLoS computational biology **5**(12), 1000584 (2009)

[174] Ma, J., Wang, S., Wang, Z., Xu, J.: Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. Bioinformatics **31**(21), 3506–3513 (2015)

[175] Eickholt, J., Cheng, J.: A study and benchmark of dncon: a method for protein residue-residue contact prediction using deep networks. In: BMC Bioinformatics, vol. 14, pp. 1–10 (2013). BioMed Central

[176] Xiong, D., Zeng, J., Gong, H.: A deep learning framework for improving long-range residue–residue contact prediction using a hierarchical strategy. Bioinformatics **33**(17), 2675–2683 (2017)

[177] Michel, M., Menéndez Hurtado, D., Elofsson, A.: PconsC4: fast, accurate and hassle-free contact predictions. Bioinformatics **35**(15), 2677–2679 (2019)

[178] Ji, S., Oruç, T., Mead, L., Rehman, M.F., Thomas, C.M., Butterworth, S., Winn, P.J.: DeepCDpred: inter-residue distance and contact prediction for improved prediction of protein structure. PloS one **14**(1), 0205214 (2019)

[179] Chen, P., Li, J.: Prediction of protein long-range contacts using an ensemble of genetic algorithm classifiers with sequence profile centers. BMC structural biology **10**(1), 1–13 (2010)

[180] Li, Y., Fang, Y., Fang, J.: Predicting residue–residue contacts using random forest models. Bioinformatics **27**(24), 3379–3384 (2011)

[181] Eickholt, J., Cheng, J.: Predicting protein residue–residue contacts using deep networks and boosting. Bioinformatics **28**(23), 3066–3072 (2012)

[182] Di Lena, P., Nagata, K., Baldi, P.: Deep architectures for protein contact map prediction. Bioinformatics **28**(19), 2449–2457 (2012)

[183] Zhang, H., Bei, Z., Xi, W., Hao, M., Ju, Z., Saravanan, K.M., Zhang, H., Guo, N., Wei, Y.: Evaluation of residue-residue contact prediction methods: From retrospective to prospective. PLOS Computational Biology **17**(5), 1009027 (2021)

[184] Billings, W.M., Morris, C.J., Della Corte, D.: The whole is greater than its parts: ensembling improves protein contact prediction. Scientific Reports **11**(1), 1–7 (2021)

[185] Jones, D.T., Buchan, D.W., Cozzetto, D., Pontil, M.: PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics **28**(2), 184–190 (2012)

[186] Kamisetty, H., Ovchinnikov, S., Baker, D.: Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. Proceedings of the National Academy of Sciences **110**(39), 15674–15679 (2013)

[187] Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., Aurell, E.: Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. Physical Review E **87**(1), 012707 (2013)

[188] He, B., Mortuza, S., Wang, Y., Shen, H.-B., Zhang, Y.: NeBcon: protein contact map prediction using neural network training coupled with naïve bayes classifiers. Bioinformatics **33**(15), 2296–2306 (2017)

[189] Wang, Z., Xu, J.: Predicting protein contact map using evolutionary and physical constraints by integer programming. Bioinformatics **29**(13), 266–273 (2013)

[190] Fukuda, H., Tomii, K.: DeepECA: an end-to-end learning framework for protein contact prediction from a multiple sequence alignment. BMC bioinformatics **21**(1), 1–15 (2020)

[191] Yang, H., Wang, M., Yu, Z., Zhao, X.-M., Li, A.: GANcon: protein contact map prediction with deep generative adversarial network. IEEE Access **8**, 80899–80907 (2020)

[192] Wu, T., Hou, J., Adhikari, B., Cheng, J.: Analysis of several key factors influencing deep learning-based inter-residue contact prediction. Bioinformatics **36**(4), 1091–1098 (2020)

[193] Wu, T., Liu, J., Guo, Z., Hou, J., Cheng, J.: MULTICOM2 open-source protein structure prediction system powered by deep learning and distance prediction. Scientific Reports **11**(1), 1–9 (2021)

[194] Jing, X., Xu, J.: Improved protein model quality assessment by integrating sequential and pairwise features using deep learning. Bioinformatics **36**(22-23), 5361–5367 (2020)

[195] Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., Baker,

D.: Improved protein structure refinement guided by deep learning based accuracy estimation. Nature communications **12**(1), 1–11 (2021)

[196] Shuvo, M.H., Bhattacharya, S., Bhattacharya, D.: QDeep: distance-based protein model quality estimation by residue-level ensemble error classifications using stacked deep residual neural networks. Bioinformatics **36**(Supplement_1), 285–291 (2020)

[197] Huang, H., Gong, X.: A review of protein inter-residue distance prediction. Current Bioinformatics **15**(8), 821–830 (2020)

[198] Ding, W., Gong, H.: Predicting the real-valued inter-residue distances for proteins. Advanced Science **7**(19), 2001314 (2020)

[199] Jain, A., Terashi, G., Kagaya, Y., Subramaniya, S.R.M.V., Christoffer, C., Kihara, D.: Analyzing effect of quadruple multiple sequence alignments on deep learning based protein inter-residue distance prediction. Scientific Reports **11**(1), 1–13 (2021)

[200] Su, H., Wang, W., Du, Z., Peng, Z., Gao, S.-H., Cheng, M.-M., Yang, J.: Improved protein structure prediction using a new multi-scale network and homologous templates. Advanced Science, 2102592 (2021)

[201] Zheng, W., Li, Y., Zhang, C., Zhou, X., Pearce, R., Bell, E.W., Huang, X., Zhang, Y.: Protein structure prediction using deep learning distance and hydrogen-bonding restraints in casp14. Proteins: Structure, Function, and Bioinformatics **89**(12), 1734–1751 (2021)

[202] Rahman, J., Newton, M., Islam, M.K.B., Sattar, A.: Enhancing protein inter-residue real distance prediction by scrutinising deep learning models. Scientific Reports **12**(1), 1–13 (2022)

[203] Rahman, J., Newton, M.H., Hasan, M.A.M., Sattar, A.: A stacked meta-ensemble for protein inter-residue distance prediction. Computers in Biology and Medicine **148**, 105824 (2022)

[204] Kandathil, S.M., Greener, J.G., Jones, D.T.: Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins: Structure, Function, and Bioinformatics **87**(12), 1092–1099 (2019)

[205] Xu, J., Wang, S.: Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins: Structure, Function, and Bioinformatics **87**(12), 1069–1081 (2019)

[206] Adhikari, B., Cheng, J.: CONFOLD2: improved contact-driven ab initio protein structure modeling. BMC bioinformatics **19**(1), 1–5 (2018)

[207] Guo, Z., Wu, T., Liu, J., Hou, J., Cheng, J.: Improving deep learning-based protein distance prediction in casp14. Bioinformatics **37**(19), 3190–3196 (2021)

[208] Chen, C., Wu, T., Guo, Z., Cheng, J.: Combination of deep neural network with attention mechanism enhances the explainability of protein contact prediction. Proteins: Structure, Function, and Bioinformatics **89**(6), 697–707 (2021)

[209] Shen, T., Wu, J., Lan, H., Zheng, L., Pei, J., Wang, S., Liu, W., Huang, J.: When homologous sequences meet structural decoys: Accurate contact prediction by tfold in casp14—(tfold for casp14 contact prediction). Proteins: Structure, Function, and Bioinformatics **89**(12), 1901–1910 (2021)

[210] Zhang, J., Zhang, Y.: A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PloS one **5**(10), 15386 (2010)

[211] Zhou, H., Skolnick, J.: GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical journal **101**(8), 2043–2052 (2011)

[212] Anishchenko, I., Baek, M., Park, H., Dauparas, J., Hiranuma, N., Mansoor, S., Humphrey, I., Baker, D.: Protein structure prediction guided by predicted inter-residue geometries. In: Fourteenth Meeting of Critical Assessment of Techniques for Protein Structure Prediction, p. 30 (2020)

[213] Zhang, C., Zhang, Y.: Protein 3D structure prediction by d-quark in CASP14. In: Fourteenth Meeting of Critical Assessment of Techniques for Protein Structure Prediction, p. 220 (2020)

[214] Li, Y., Zheng, W., Zhang, C., Bell, E., Huang, X., Pearce, R., Zhou, X., Zhang, Y.: Protein 3D structure prediction by DI-TASSER in CASP14. CASP **14**, 339–341 (2020)

[215] Cheng, J., Tegge, A.N., Baldi, P.: Machine learning methods for protein structure prediction. IEEE reviews in biomedical engineering **1**, 41–49 (2008)

[216] Chuang, C.-C., Chen, C.-Y., Yang, J.-M., Lyu, P.-C., Hwang, J.-K.: Relationship between protein structures and disulfide-bonding patterns. Proteins: Structure, Function, and Bioinformatics: Structure, Function, and Bioinformatics **53**(1), 1–5 (2003)

[217] Lin, H.-H., Tseng, L.-Y.: DBCP: a web server for disulfide bonding connectivity pattern prediction without the prior knowledge of the bonding

state of cysteines. Nucleic acids research **38**(suppl_2), 503–507 (2010)

[218] Yang, J., He, B.-J., Jang, R., Zhang, Y., Shen, H.-B.: Accurate disulfide-bonding network predictions improve ab initio structure prediction of cysteine-rich proteins. Bioinformatics **31**(23), 3773–3781 (2015)

[219] Liu, Z.-L., Hu, J.-H., Jiang, F., Wu, Y.-D.: CRiSP: accurate structure prediction of disulfide-rich peptides with cystine-specific sequence alignment and machine learning. Bioinformatics **36**(11), 3385–3392 (2020)

[220] Mishra, A., Kabir, M.W.U., Hoque, M.T.: diSBPred: a machine learning based approach for disulfide bond prediction. Computational Biology and Chemistry **91**, 107436 (2021)

[221] Niu, S., Huang, T., Feng, K.-Y., He, Z., Cui, W., Gu, L., Li, H., Cai, Y.-D., Li, Y.: Inter-and intra-chain disulfide bond prediction based on optimal feature selection. Protein and peptide letters **20**(3), 324–335 (2013)

[222] Rodriguez, C., Chowriappa, P., Dua, S., *et al.*: Local similarity matrix for cysteine disulfide connectivity prediction from protein sequences. IEEE/ACM transactions on computational biology and bioinformatics **17**(4), 1276–1289 (2019)

[223] Cheng, J., Baldi, P.: Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. Bioinformatics **21**(suppl_1), 75–84 (2005)

[224] Kalisman, N., Levi, A., Maximova, T., Reshef, D., Zafriri-Lynn, S., Gleyzer, Y., Keasar, C.: MESHI: a new library of Java classes for molecular modeling. Bioinformatics **21**(20), 3931–3932 (2005)

[225] Li, Y., Roy, A., Zhang, Y.: HAAD: a quick algorithm for accurate prediction of hydrogen atoms in protein structures. PloS one **4**(8), 6701 (2009)

[226] Bhattacharya, D., Cheng, J.: 3Drefine: consistent protein structure refinement by optimizing hydrogen bonding network and atomic-level energy minimization. Proteins: Structure, Function, and Bioinformatics **81**(1), 119–131 (2013)

[227] Bhattacharya, D., Nowotny, J., Cao, R., Cheng, J.: 3Drefine: an interactive web server for efficient protein structure refinement. Nucleic acids research **44**(W1), 406–409 (2016)

[228] Dou, J., Vorobieva, A.A., Sheffler, W., Doyle, L.A., Park, H., Bick, M.J.,

Mao, B., Foight, G.W., Lee, M.Y., Gagnon, L.A., *et al.*: De novo design of a fluorescence-activating β-barrel. Nature **561**(7724), 485–491 (2018)

[229] Li, Y., Zhang, C., Zheng, W., Zhou, X., Bell, E., Yu, D., Zhang, Y.: Learning deep statistical potentials for protein folding. CASP **14**, 72–73 (2020)

[230] Kryshtafovych, A., Monastyrskyy, B., Fidelis, K., Schwede, T., Tramontano, A.: Assessment of model accuracy estimations in CASP12. Proteins: Structure, Function, and Bioinformatics **86**, 345–360 (2018)

[231] Won, J., Baek, M., Monastyrskyy, B., Kryshtafovych, A., Seok, C.: Assessment of protein model structure accuracy estimation in CASP13: challenges in the era of deep learning. Proteins: Structure, Function, and Bioinformatics **87**(12), 1351–1360 (2019)

[232] Karasikov, M., Pagès, G., Grudinin, S.: Smooth orientation-dependent scoring function for coarse-grained protein quality assessment. Bioinformatics **35**(16), 2801–2808 (2019)

[233] Cao, R., Adhikari, B., Bhattacharya, D., Sun, M., Hou, J., Cheng, J.: QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. Bioinformatics **33**(4), 586–588 (2017)

[234] Maghrabi, A.H., McGuffin, L.J.: ModFOLD6: an accurate web server for the global and local quality estimation of 3D protein models. Nucleic acids research **45**(W1), 416–421 (2017)

[235] Benkert, P., Tosatto, S.C., Schomburg, D.: QMEAN: a comprehensive scoring function for model quality assessment. Proteins: Structure, Function, and Bioinformatics **71**(1), 261–277 (2008)

[236] Lee, G.R., Won, J., Heo, L., Seok, C.: GalaxyRefine2: simultaneous refinement of inaccurate local regions and overall protein structure. Nucleic acids research **47**(W1), 451–455 (2019)

[237] Bhattacharya, D.: refineD: improved protein structure refinement using machine learning based restrained relaxation. Bioinformatics **35**(18), 3320–3328 (2019)

[238] MacCarthy, E., Perry, D., Kc, D.B.: Advances in protein super-secondary structure prediction and application to protein structure prediction. Methods in molecular biology (Clifton, NJ) **1958**, 15–45 (2019)

[239] Zou, D., He, Z., He, J., Xia, Y.: Supersecondary structure prediction using Chou's pseudo amino acid composition. Journal of Computational

Chemistry **32**(2), 271–278 (2011)

[240] Li, C., Wang, X.-F., Chen, Z., Zhang, Z., Song, J.: Computational characterization of parallel dimeric and trimeric coiled-coils using effective amino acid indices. Molecular BioSystems **11**(2), 354–360 (2015)

[241] Kou, G., Feng, Y.: Identify five kinds of simple super-secondary structures with quadratic discriminant algorithm based on the chemical shifts. Journal of theoretical biology **380**, 392–398 (2015)

[242] Wang, X., Zhou, Y., Yan, R.: AAFreqCoil: a new classifier to distinguish parallel dimeric and trimeric coiled coils. Molecular BioSystems **11**(7), 1794–1801 (2015)

[243] Wood, C.W., Woolfson, D.N.: CCBuilder2.0: powerful and accessible coiled-coil modeling. Protein Science **27**(1), 103–111 (2018)

[244] Flot, M., Mishra, A., Kuchi, A.S., Hoque, M.T.: StackSSSPred: a stacking-based prediction of supersecondary structure from sequence. Methods in molecular biology (Clifton, NJ) **1958**, 101–122 (2019)

[245] Hu, X.-z., Long, H.-x., Ding, C.-j., Gao, S.-j., Hou, R.: Using random forest algorithm to predict super-secondary structure in proteins. The Journal of Supercomputing **76**(5), 3199–3210 (2020)

[246] Chen, K., Kurgan, L.: Computational prediction of secondary and supersecondary structures. In: Protein Supersecondary Structures, pp. 63–86. Springer, Switzerland (2012)

[247] Scott, W.R., Hünenberger, P.H., Tironi, I.G., Mark, A.E., Billeter, S.R., Fennen, J., Torda, A.E., Huber, T., Krüger, P., van Gunsteren, W.F.: The GROMOS biomolecular simulation program package. The Journal of Physical Chemistry A **103**(19), 3596–3607 (1999)

[248] Damm, W., Frontera, A., Tirado–Rives, J., Jorgensen, W.L.: OPLS all-atom force field for carbohydrates. Journal of computational chemistry **18**(16), 1955–1970 (1997)

[249] Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., *et al.*: The rosetta all-atom energy function for macromolecular modeling and design. Journal of chemical theory and computation **13**(6), 3031–3048 (2017)

[250] Jones, D.T., McGuffin, L.J.: Assembling novel protein folds from super-secondary structural fragments. Proteins: Structure, Function, and Bioinformatics **53**(S6), 480–485 (2003)

[251] O'Meara, M.J., Leaver-Fay, A., Tyka, M.D., Stein, A., Houlihan, K., DiMaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., *et al.*: Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with rosetta. Journal of chemical theory and computation **11**(2), 609–622 (2015)

[252] Venske, S.M., Gonçalves, R.A., Benelli, E.M., Delgado, M.R.: ADEMO/D: an adaptive differential evolution for protein structure prediction problem. Expert Systems with Applications **56**, 209–226 (2016)

[253] Shuid, A.N., Kempster, R., McGuffin, L.J.: ReFOLD: a server for the refinement of 3D protein models guided by accurate quality estimates. Nucleic acids research **45**(W1), 422–428 (2017)

[254] Gao, S., Song, S., Cheng, J., Todo, Y., Zhou, M.: Incorporation of solvent effect into multi-objective evolutionary algorithm for improved protein structure prediction. IEEE/ACM transactions on computational biology and bioinformatics **15**(4), 1365–1378 (2017)

[255] Ramyachitra, D., Ajeeth, A.: MODCSA-CA: a multi objective diversity controlled self adaptive cuckoo algorithm for protein structure prediction. Gene Reports **8**, 100–106 (2017)

[256] Song, S., Gao, S., Chen, X., Jia, D., Qian, X., Todo, Y.: Aimoes: Archive information assisted multi-objective evolutionary strategy for ab initio protein structure prediction. Knowledge-Based Systems **146**, 58–72 (2018)

[257] Song, S., Ji, J., Chen, X., Gao, S., Tang, Z., Todo, Y.: Adoption of an improved pso to explore a compound multi-objective energy function in protein structure prediction. Applied Soft Computing **72**, 539–551 (2018)

[258] Yang, Y., Zhou, Y.: Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins: Structure, Function, and Bioinformatics **72**(2), 793–803 (2008)

[259] Zaman, A.B., Shehu, A.: Balancing multiple objectives in conformation sampling to control decoy diversity in template-free protein structure prediction. BMC bioinformatics **20**(1), 1–17 (2019)

[260] Varela, D., Santos, J.: Niching methods integrated with a differential evolution memetic algorithm for protein structure prediction. Elsevier (2022)

[261] Dill, K.A.: Theory for the folding and stability of globular proteins. Biochemistry **24**(6), 1501–1509 (1985)

[262] Miyazawa, S., Jernigan, R.L.: Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules **18**(3), 534–552 (1985)

[263] Miyazawa, S., Jernigan, R.L.: Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. Journal of molecular biology **256**(3), 623–644 (1996)

[264] Adhikari, B., Bhattacharya, D., Cao, R., Cheng, J.: CONFOLD: residue-residue contact-guided ab initio protein folding. Proteins: Structure, Function, and Bioinformatics **83**(8), 1436–1449 (2015)

[265] Liu, J., Zhou, X.-G., Zhang, Y., Zhang, G.-J.: CGLFold: a contact-assisted de novo protein structure prediction using global exploration and loop perturbation sampling algorithm. Bioinformatics **36**(8), 2443–2450 (2020)

[266] Cai, Y., Li, X., Sun, Z., Lu, Y., Zhao, H., Hanson, J., Paliwal, K., Litfin, T., Zhou, Y., Yang, Y.: SPOT-Fold: fragment-free protein structure prediction guided by predicted backbone structure and contact map. Journal of Computational Chemistry **41**(8), 745–750 (2020)

[267] Chen, X., Song, S., Ji, J., Tang, Z., Todo, Y.: Incorporating a multi-objective knowledge-based energy function into differential evolution for protein structure prediction. Information Sciences **540**, 69–88 (2020)

[268] Mishra, A., Iqbal, S., Hoque, M.T.: Discriminate protein decoys from native by using a scoring function based on ubiquitous phi and psi angles computed for all atom. Journal of theoretical biology **398**, 112–121 (2016)

[269] Mabrouk, M., Werner, T., Schneider, T., Putz, I., Brock, O.: Analysis of free modelling predictions by RBO aleph in CASP11. Proteins **84**, 87–104 (2015)

[270] Bhattacharya, D., Cao, C. Renzhi, Jianlin: UniCon3D: de novo protein structure prediction using united-residue conformational search via stepwise, probabilistic sampling. Bioinformatics **32**(18), 2791–2799 (2016)

[271] de Lima Corrêa, L., Borguesan, B., Krause, M.J., Dorn, M.: Three-dimensional protein structure prediction based on memetic algorithms. Computers & Operations Research **91**, 160–177 (2018)

[272] Brünger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu,

N.S., *et al.*: Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallographica Section D: Biological Crystallography **54**(5), 905–921 (1998)

[273] Ovchinnikov, S., Park, H., Kim, D.E., DiMaio, F., Baker, D.: Protein structure prediction using rosetta in casp12. Proteins: Structure, Function, and Bioinformatics **86**, 113–121 (2018)

[274] Heo, L., Feig, M.: High-accuracy protein structures by combining machine-learning with physics-based refinement. Proteins: Structure, Function, and Bioinformatics **88**(5), 637–642 (2020)

[275] Brunger, A.T.: Version 1.2 of the crystallography and nmr system. Nature protocols **2**(11), 2728–2733 (2007)

[276] Simons, K.T., Kooperberg, C., Huang, E., Baker, D.: Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. Journal of molecular biology **268**(1), 209–225 (1997)

[277] Gront, D., Kulp, D.W., Vernon, R.M., Strauss, C.E., Baker, D.: Generalized fragment picking in rosetta: design, protocols and applications. PloS one **6**(8), 23294 (2011)

[278] de Oliveira, S.H., Shi, J., Deane, C.M.: Building a better fragment library for de novo protein structure prediction. PloS one **10**(4), 0123998 (2015)

[279] Santos, K.B., Trevizani, R., Custódio, F.L., Dardenne, L.E.: Profrager web server: Fragment libraries generation for protein structure prediction. In: Proceedings of the International Conference on Bioinformatics & Computational Biology (BIOCOMP), p. 38 (2015). The Steering Committee of The World Congress in Computer Science, Computer ...

[280] Bhattacharya, D., Adhikari, B., Li, J., Cheng, J.: Fragsion: ultra-fast protein fragment library generation by iohmm sampling. Bioinformatics **32**(13), 2059–2061 (2016)

[281] Wang, T., Yang, Y., Zhou, Y., Gong, H.: Lrfraglib: an effective algorithm to identify fragments for de novo protein structure prediction. Bioinformatics **33**(5), 677–684 (2017)

[282] Wang, T., Qiao, Y., Ding, W., Mao, W., Zhou, Y., Gong, H.: Improved fragment sampling for ab initio protein structure prediction using deep neural networks. Nature Machine Intelligence **1**(8), 347–355 (2019)

[283] Chaudhury, S., Lyskov, S., Gray, J.J.: PyRosetta: a script-based interface for implementing molecular modeling algorithms using rosetta.

Bioinformatics **26**(5), 689–691 (2010)

[284] Leaver-Fay, A., Tyka, M., Lewis, S.M., Lange, O.F., Thompson, J., Jacak, R., Kaufman, K.W., Renfrew, P.D., Smith, C.A., Sheffler, W., *et al.*: ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. In: Methods in Enzymology vol. 487, pp. 545–574. Elsevier, Amsterdam, Netherlands (2011)

[285] Liu, S., Wang, T., Xu, Q., Shao, B., Yin, J., Liu, T.-Y.: Complementing sequence-derived features with structural information extracted from fragment libraries for protein structure prediction. BMC bioinformatics **22**(1), 1–18 (2021)

[286] Mortuza, S., Zheng, W., Zhang, C., Li, Y., Pearce, R., Zhang, Y.: Improving fragment-based ab initio protein structure assembly using low-accuracy contact-map predictions. Nature communications **12**(1), 1–12 (2021)

[287] Colbes, J., Corona, R.I., Lezcano, C., Rodríguez, D., Brizuela, C.A.: Protein side-chain packing problem: is there still room for improvement? Briefings in bioinformatics **18**(6), 1033–1043 (2017)

[288] Huang, X., Pearce, R., Zhang, Y.: FASPR: an open-source tool for fast and accurate protein side-chain packing. Bioinformatics **36**(12), 3758–3765 (2020)

[289] Xu, G., Wang, Q., Ma, J.: OPUS-Rota3: Improving protein side-chain modeling by deep neural networks and ensemble methods. Journal of Chemical Information and Modeling **60**(12), 6691–6697 (2020)

[290] Maguire, J.B., Haddox, H.K., Strickland, D., Halabiya, S.F., Coventry, B., Griffin, J.R., Pulavarti, S.V.K., Cummins, M., Thieker, D.F., Klavins, E., *et al.*: Perturbing the energy landscape for improved packing during computational protein design. Proteins: Structure, Function, and Bioinformatics **89**(4), 436–449 (2021)

[291] Cao, Y., Song, L., Miao, Z., Hu, Y., Tian, L., Jiang, T.: Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation. Bioinformatics **27**(6), 785–790 (2011)

[292] Miao, Z., Cao, Y., Jiang, T.: RASP: rapid modeling of protein side chain conformations. Bioinformatics **27**(22), 3117–3122 (2011)

[293] Jumper, J.M., Faruk, N.F., Freed, K.F., Sosnick, T.R.: Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. PLoS computational biology **14**(12), 1006342 (2018)

[294] Nagata, K., Randall, A., Baldi, P.: SIDEpro: a novel machine learning approach for the fast and accurate prediction of side-chain conformations. Proteins: Structure, Function, and Bioinformatics **80**(1), 142–153 (2012)

[295] Kingsford, C.L., Chazelle, B., Singh, M.: Solving and analyzing side-chain positioning problems using linear and integer programming. Bioinformatics **21**(7), 1028–1039 (2005)

[296] Huang, X., Han, K., Zhu, Y.: Systematic optimization model and algorithm for binding sequence selection in computational enzyme design. Protein Science **22**(7), 929–941 (2013)

[297] Gordon, D.B., Mayo, S.L.: Branch-and-terminate: a combinatorial optimization algorithm for protein design. Structure **7**(9), 1089–1098 (1999)

[298] Canutescu, A.A., Shelenkov, A.A., Dunbrack Jr, R.L.: A graph-theory algorithm for rapid protein side-chain prediction. Protein science **12**(9), 2001–2014 (2003)

[299] Xu, J., Berger, B.: Fast and accurate algorithms for protein side-chain packing. Journal of the ACM (JACM) **53**(4), 533–557 (2006)

[300] Peterson, R.W., Dutton, P.L., Wand, A.J.: Improved side-chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Science **13**(3), 735–751 (2004)

[301] Liu, Z., Jiang, L., Gao, Y., Liang, S., Chen, H., Han, Y., Lai, L.: Beyond the rotamer library: Genetic algorithm combined with the disturbing mutation process for upbuilding protein side-chains. Proteins: Structure, Function, and Bioinformatics **50**(1), 49–62 (2003)

[302] Xu, G., Ma, T., Du, J., Wang, Q., Ma, J.: OPUS-Rota2: an improved fast and accurate side-chain modeling method. Journal of chemical theory and computation **15**(9), 5154–5160 (2019)

[303] Varela, D., Santos, J.: A protein folding model using the face-centered cubic lattice model. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion, pp. 1674–1678 (2017)

[304] Irbäck, A., Peterson, C., Potthast, F., Sommelius, O.: Local interactions and protein folding: A three-dimensional off-lattice approach. The Journal of chemical physics **107**(1), 273–282 (1997)

[305] Chan, T., Jankovic, B., Le, V., Navernioukk, I.: Comparative Study of Hydrophobic-Polar and Miyazawa-Jernigan Energy Functions in Protein Folding on a Cubic Lattice Using Pruned-Enriched Rosenbluth Monte

Carlo Algorithm (2004)

[306] Chi, P.B., Kim, D., Lai, J.K., Bykova, N., Weber, C.C., Kubelka, J., Liberles, D.A.: A new parameter-rich structure-aware mechanistic model for amino acid substitution during evolution. Proteins: Structure, Function, and Bioinformatics **86**(2), 218–228 (2018)

[307] Berrera, M., Molinari, H., Fogolari, F.: Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC bioinformatics **4**(1), 1–26 (2003)

[308] Rashid, M.A., Shatabda, S., Newton, M.H., Hoque, M.T., Pham, D.N., Sattar, A.: Random-walk: a stagnation recovery technique for simplified protein structure prediction. In: Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, pp. 620–622 (2012)

[309] Rashid, M.A., Newton, M.H., Hoque, M.T., Sattar, A.: A local search embedded genetic algorithm for simplified protein structure prediction. In: 2013 IEEE Congress on Evolutionary Computation, pp. 1091–1098 (2013). IEEE

[310] Rashid, M.A., Newton, M., Hoque, M., Sattar, A., et al.: Mixing energy models in genetic algorithms for on-lattice protein structure prediction. BioMed research international **2013** (2013)

[311] Rashid, M.A., Khatib, F., Hoque, M.T., Sattar, A.: An enhanced genetic algorithm for ab initio protein structure prediction. IEEE Transactions on Evolutionary Computation **20**(4), 627–644 (2015)

[312] Do Duc, D., Dinh, P.T., Anh, V.T.N., Linh-Trung, N.: An efficient ant colony optimization algorithm for protein structure prediction. In: 2018 12th International Symposium on Medical Information and Communication Technology (ISMICT), pp. 1–6 (2018). IEEE

[313] Takahashi, T., Chikenji, G., Tokita, K.: Lattice protein design using bayesian learning. Physical Review E **104**(1), 014404 (2021)

[314] Atari, M., Majd, N.: 2D HP protein folding using quantum genetic algorithm. In: 2022 27th International Computer Conference, Computer Society of Iran (CSICC), pp. 1–8 (2022). IEEE

[315] Zhang, L., Ma, H., Qian, W., Li, H.: Protein structure optimization using improved simulated annealing algorithm on a three-dimensional ab off-lattice model. Computational Biology and Chemistry **85**, 107237 (2020)

[316] Rakhshani, H., Idoumghar, L., Ghambari, S., Lepagnot, J., Brévilliers, M.: On the performance of deep learning for numerical optimization: an application to protein structure prediction. Applied Soft Computing **110**, 107596 (2021)

[317] Xia, Y.-H., Peng, C.-X., Zhou, X.-G., Zhang, G.-J.: A sequential niche multimodal conformational sampling algorithm for protein structure prediction. Bioinformatics **37**(23), 4357–4365 (2021)

[318] Shuchun, Y., Xianxiang, L., Xue, T., Ming, P.: Protein structure prediction based on particle swarm optimization and tabu search strategy. BMC bioinformatics **23**(10), 1–10 (2022)

[319] Shatabda, S., Newton, M., Rashid, M.A., Pham, D.N., Sattar, A.: How good are simplified models for protein structure prediction? Advances in bioinformatics **2014** (2014)

[320] Shatabda, S., Newton, M.H., Sattar, A.: Simplified lattice models for protein structure prediction: how good are they? In: Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)

[321] Steinegger, M., Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature biotechnology **35**(11), 1026–1028 (2017)

M.M. Mohamed Mufassirin. M.M. Mohamed Mufassirin is pursuing a PhD degree at Griffith University, Australia. Currently, He is a Lecturer in Computer Science at the South Eastern University of Sri Lanka. His research interests include Artificial Intelligence, Bioinformatics, Machine Learning and Protein Design. He obtained his M.Sc in Computer Science at the University of Peradeniya and B.Sc (Hons) in Computer Science at the South Eastern University of Sri Lanka.

M. A. Hakim Newton. M.A.H. Newton is a lecturer in Data Science in the School of Information and Physical Sciences at the University of Newcastle, Australia. He obtained his PhD from Strathclyde University, United Kingdom and his MScEngg and BScEngg from Bangladesh University of Engineering and Technology (BUET). Dr Newton was a researcher in the National ICT Australia (NICTA) and in the Institute for Integrated and Intelligent Systems (IIIS) and the School of ICT at Griffith University, Australia. His research interests are in Intelligent Search, Machine Learning, and Bioinformatics.

A. Sattar. A. Sattar is a professor at the School of ICT, Griffith University, Australia. He was the founding Director of the Institute for Integrated and Intelligent Systems at Griffith. He was also the Education Director at Queensland Research Lab (QRL) at National ICT Australia (NICTA). He won a number of ARC discovery grants and international awards for his work in Artificial Intelligence.