IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 23, 2022, accepted May 15, 2022, date of publication May 23, 2022, date of current version May 27, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3176945

Constraint Guided Neighbor Generation for Protein Structure Prediction

RIANON ZAMAN[1], M. A. HAKIM NEWTON[2,3], FERESHTEH MATAEIMOGHADAM[1], AND ABDUL SATTAR[1,3]

[1]School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia
[2]School of Information and Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia
[3]Institute for Integrated and Intelligent Systems, Griffith University, Nathan, QLD 4111, Australia

Corresponding authors: Rianon Zaman (rianon.zaman@griffithuni.edu.au) and M. A. Hakim Newton (mahakim.newton@newcastle.edu.au)

This work was supported in part by the Australian Research Council Discovery under Grant DP180102727.

ABSTRACT Protein structure prediction (PSP) is essential for drug discovery. PSP involves minimising an unknown scoring function over an astronomical search space. PSP has achieved significant progress recently via end-to-end deep learning models that require enormous computational resources and almost all known proteins as training data. In this paper, we develop a conformational search method for PSP based on scoring functions involving geometric constraints learnt by deep learning models. When machine learning models achieve generality and thus obviously loose accuracy, conformational search methods could perform protein-specific fine tuning of the predicted conformations. However, effective conformational sampling in PSP remains a key challenge. Existing conformational search algorithms adopt random selection approaches for neighbor generation and thus greatly depend on luck. We propose a new approach to analyse geometric constraint-based scores, to identify the regions of the current conformations causing inferior scores, and to alter the identified regions to generate neighbour conformations. Our approach prefers informed decisions to random selections from an artificial intelligence perspective. The proposed method also provides promising search guidance as it obtains significant improvements from given initial conformations. On a set of 35 benchmark proteins of varying types and sizes, our algorithm significantly outperforms state-of-the-art PSP search algorithms that use random sampling with a similar scoring function: the improvement is about 1Å better average in root mean square deviation (RMSD) values. Our sample generation approach could be used in other bioinformatics research areas requiring search.

INDEX TERMS Protein structure prediction, search-based optimisation, neighbour generation.

I. INTRODUCTION

Proteins are sequences of amino acid (AA) residues. Proteins fold into three dimensional structures. A protein's AA sequence essentially determines its native structure having the lowest free energy and the native structure essentially determines its function. By docking on a disease protein's native structure, drug molecules inhibit its functions. Protein structure prediction (PSP) by in vitro methods are time consuming, costlier, and failure prone. Computational PSP approaches minimise unknown scoring functions over astronomical search spaces and find decoy structures.

The associate editor coordinating the review of this manuscript and approving it for publication was Abdullah Iliyasu.

PSP has achieved significant progress recently via AlphaFold2's end-to-end deep learning models [1]. However, AlphaFold2 needs enormous computational resources and uses almost all known and unknown proteins in its training. Moreover, AlphaFold2's algorithmic details are not open. Although its trained model is available, because of the computational resource requirement, most research labs cannot run it locally. Its google Collab interface provides only a restricted access to AlphaFold2.

After AlphaFold2, for further scientific advancement, the immediate challenge to the PSP community is to obtain at least AlphaFold2's accuracy level but using simpler and more efficient PSP methods that depend on fewer training proteins. Further, alternative methods that are based on conformational search approaches could also be investigated.

Furthermore, PSP methods should be made available. In this paper, we investigate conformational search methods for PSP using proxy energy functions i.e. scoring functions based on geometric constraints learnt by deep learning models [2]–[4]. PSP search methods include Monte Carlo algorithms [5], evolutionary algorithms [6], multi-objective optimisation [7], sequential search [8], differential evolution [9], memetic algorithms [10], [11], and gradient descent algorithms [2], [4], [12]. In general, these iterative search algorithms generate neighbour conformations i.e. three dimensional protein structures randomly from the current conformations, evaluate the generated neighbour conformations using chosen scoring functions, and select the best neighbour conformations as the current conformations for the next iterations. As such the conformation evaluation phase only indirectly guides the search while the conformation generation phase largely remains unguided and dependent on luck.

From artificial intelligence perspectives, our motivation is to generate neighbour conformations based on informed decisions. So we detect problematic parts of the current conformations and make changes mainly in those identified parts. To detect the problematic parts, we use a constraint-guided approach that helps analyse unsatisfied geometric constraints causing inferior scores of the current conformations. Our approach is simple and it explains the selection decisions made by the neighbourhood generation procedure. To the best of our knowledge, this is the first attempt in taking an informed approach in neighbour generation for PSP. The proposed strategy could be useful in other bioinformatics search problems that include structure based drug design.

We evaluate our constraint guided neighbour generation approach within a simple local search framework. For protein structure representation, we use dihedral angles but also compute Cartesian coordinates of the atoms. Moreover, we consider only the main chains or the backbones of the protein structures and exclude the side chains of the amino acid residues. For protein structure evaluation, we use scoring functions based on predicted residue-residue distances. Our algorithm has been implemented on our newly developed constraint-based PSP search platform Koala. We use a set of benchmark proteins of varying types and sizes. Experimental results show that our constraint-based neighbour generation approach significantly outperforms other random-based PSP search approaches. Our proposed approach also significantly outperforms state-of-the-art PSP search algorithms that use random sampling with similar scoring functions.

The rest of the paper explores the related work, details of the problem formulation, illustrates the main idea, describes the implementation details, and presents our experimental results and analyses, and also presents our conclusions.

II. RELATED WORK

Considering the relevance with this work, we mainly explore the search and optimisation approach for PSP.

The free energy of a protein has not been precisely known or defined so far, but physical (Van der Walls forces), chemical (bond energies), and electrostatic (Coulomb forces) energy components have been used in protein structure scoring functions based on molecular dynamics e.g. in CHARMM [13]. Another such successful scoring function used in PSP research is the ROSETTA [14] energy function. Nevertheless, energy functions that involve all-atomic details are computationally very expensive. Note that the energy value is to be computed for each conformation generated during search.

Quark [5] constructs structures using fragment assembly, refines them using replica-exchange Monte Carlo simulations, and uses a composite knowledge-based force field. Quark's force field has eleven terms that include atomic-level, residue-level, and topology-level terms. These terms are knowledge based but also have direct physical basis.

Differential evolution (DE) has been very effective in PSP. An underestimation-assisted global and local cooperative DE (GLCDE) improves the search capability of DE [6]. In GLCDE, the global phase tries to locate promising regions quickly whereas the local phase serves as a local search for improving convergence. To get the underestimation of the objective function, on the basis of the abstract convexity theory, GLCDE designs an adaptive underestimation model in which the slope control factor of the supporting vectors is dynamically updated based on the evaluated trial individual. AIMOES [7] is a multi-objective optimisation technique, which reuses past search experiences carried by a decision maker to select representative solutions. It includes three different physical energy terms: bond energy, non-bond energy, and solvent accessible surface area. MODE-K [9] presents a multi-objective differential evolution algorithm and maintains an archive of optimal solutions. MODE-K uses RWplus [15] as the energy function and decomposes the energy function into two terms to get multiple objectives: a distance-dependent energy term and an orientation-dependent term.

SAINT2 [8] uses sequential search along with an independent fragment-assembly approach to predict both sequential or non-sequential structures. SAINT2 uses a combination of knowledge based and physical potentials as energy functions.

PSP search methods also include memetic algorithms. A knowledge based memetic algorithm [11] shows the angle Probability List strategy is quite useful in identifying distinct structural patterns. As an energy function, it uses ROSETTA [14] and solvent accessible surface area (SASA) [16].

Recently, trRosetta [2], [12] claims that gradient descent algorithm is useful in solving PSP problems. As energy function, trRosetta uses ROSETTA [14] energy function.

As alternatives to scoring functions based on molecular dynamics, knowledge based scoring functions obtained by machine learning algorithms have also been used in PSP search e.g. residue-residue *distance maps* and residue-residue contact maps (whether residue-residue distances are within

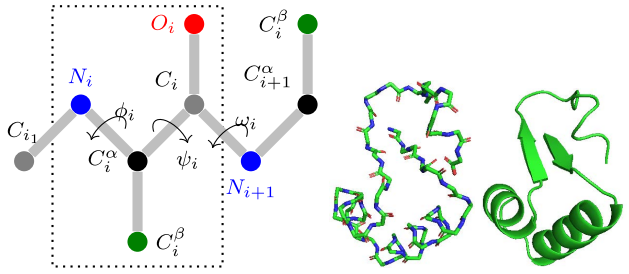


FIGURE 1. Left: an amino acid (dotted boxed) and protein backbone angles ϕ, ψ, and ω. Middle: a protein's full backbone folded in a three dimensional shape. Right: protein secondary structures i.e. helixes, sheets, and loops when shown as cartoons within the three dimensional structure.

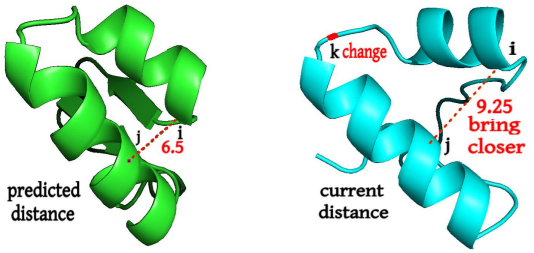


FIGURE 2. Left: distance between two helix residues *i* and *j* as predicted by a machine learning algorithm. Right: changing the dihedral angles of a loop residue *k* changes the distance between *i* and *j* to obtain the predicted distance.

8Å). In distance and contact maps, residues are represented by C^β atoms except by C^α for Glycine. SPOT-Contact [17] is a recent contact map prediction method and CONFOLD [18], [19], MULTICOM [20], and CGLFOLD [3] are recent methods that use contact maps in PSP search. Recently residue-residue distance map based scoring functions have shown promise [21], [22]. Consequently, RaptorX [23]–[25] and AlphaFold [26] predict distance maps and use them in their search algorithms for protein structures.

III. PROBLEM FORMULATION

PSP search starts with a protein's given AA sequence. FIGURE 1 Left shows AAs comprise N, C^α, C, O, and C^β atoms among others. C^α atoms are central to AAs. AAs are of 20 types and can appear any number of times in any order in a protein. Each instance of an AA in a protein is a residue. One residue's C atom is connected with another residue's N atom to form a peptide bond. Thus, we get the main chain or the backbone of a protein. Besides the main chain, each AA except Glycine has a unique side chain starting from the C^α atom and C^β is the first atom in a side chain. Assuming standard bond distances and angles, the main chain of a protein can be represented by three rotatable dihedral angles ϕ, ψ, and ω that allow folding. These three angles are respectively defined by each four successive atoms from the sequence C_{i-1}, N_i, C_i^α, C_i, N_{i+1}, C_{i+1}^α. For most proteins, ω is 180° [27], but ϕ and ψ can take any values from $-180°$ to $+180°$. The side chains at individual AAs have their own dihedral angles, but in this work, we mainly focus on searching for backbone ϕ and ψ angles of the main chain. Using backbone ϕ and ψ angles found, one can first obtain the main chain and then can later deal with the side chains to get the full protein structure.

FIGURE 1 Middle shows the backbone of an entire protein when folded into a three dimensional shape. FIGURE 1 Right shows protein structures exhibit certain local flexible and rigid regions that are called secondary structures (SS). Rigid regions such as helices and sheets are comparatively easier to be modelled since most residues in these regions have been observed to take ϕ and ψ values from very narrow ranges of about 20°. Finding the ϕ and ψ values for the flexible loop regions is challenging since they can take any fractional values in $[-180, +180]$. About 40% residues in a protein are in loops [28] and loop sampling methods strive to find dihedral angles to model them.

The energy function of a protein is not known precisely. Physical (Van der Walls forces), chemical (bond energies), and electrostatic (Coulomb forces) components are used in scoring functions based on molecular dynamics e.g. CHARMM [13] and ROSETTA [14]. These scoring functions involve all atomic-details and are computationally very expensive. Note that the scoring function is to be computed for each conformation are generated during search. As alternatives to scoring functions based on molecular dynamics, knowledge based scoring functions obtained by machine learning algorithms have been used in PSP search e.g. residue-residue distance maps [4], [26], contact maps [3], [17], [20], and angle orientations [2]. In any of these residue-residue maps, residues are represented by C^β atoms except by C^α for Glycine. Further, in contact maps, contacts denote whether residue-residue distances are within 8Å.

To evaluate our constraint-guided neighbour generation approach in PSP, in this work, we perform loop sampling with an residue-residue distance or contact map based scoring function. However, our neighbour generation approach could also be used to refine models for helixes and sheets.

IV. IDEA ILLUSTRATION

Assume i and j be two residues of a protein and d_{ij} be the prediction made by a given machine learning algorithm about the distance between residues i and j in the native structure of the protein. Also, assume c be the current conformation of the given protein during search and d_{ij}^c be the current distance between residues i and j in the conformation c. FIGURE 2 Left shows $d_{ij} = 6.5$ acts as a constraint and FIGURE 2 Right shows $d_{ij}^c = 9.25$ violates the constraint. To bring residue i and j closer to each other and thus to satisfy the constraint, we change the dihedral angles of a residue k, which is in between residues i and j.

During search, in each iteration, we heuristically choose residues i and j with d_{ij}^c the furthest from d_{ij}. Moreover, we randomly choose k from loop residues only. Note that we perform loop sampling in this work, leaving helices and sheets the same after first construction in initialisation.

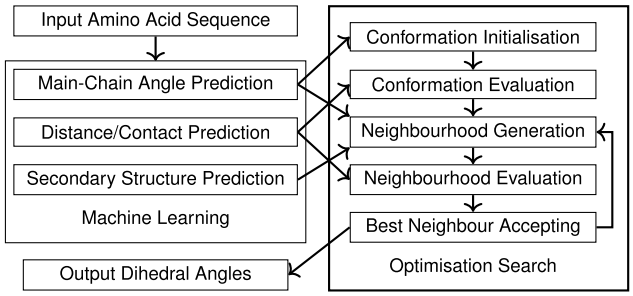


FIGURE 3. Our PSP pipeline.

Our key contribution in this paper is selection of k explicitly based on its potential to change the distance between residues i and j, which are in violation of the predicted distance constraint. Existing algorithms basically randomly select residue k without having any explicit knowledge of i and j and hence could waste the search effort.

V. IMPLEMENTATION DETAILS

FIGURE 3 shows our PSP pipeline. We start from the AA sequence of a protein. From the AA sequence of a protein, we first use machine learning approaches to predict the main-chain angles, residue-residue distance or contact maps, and secondary structures. Using the predictions in various ways, we then adopt optimisation search approaches to perform conformation initialisation and evaluation, and then in an iterative fashion generate and evaluate conformation neighbourhood and accept the best neighbour conformation as the current conformation for the next iteration. We describe each stage of our pipeline. However, our main contribution is in the optimisation search approach, more particularly in the neighbourhood generation step.

A. USING MACHINE LEARNING ALGORITHMS

There are several main-chain angle prediction methods i.e. SPIDER2 [29] SPOT-1D [2], SAP [30]. SPIDER2 predicts with mean-absolute errors (MAE) of about 19.7 and 30.3 degrees for ϕ and ψ angles respectively. For SAP, MAE values are 15.66 and 18.59 degrees respectively and for SPOT-1D, respectively 16 and 23 degrees. After some preliminary experiments, we have found that SPOT-1D predictions led to better three dimensional structures, mainly because it captures better overall shape than local structures. SPOT-1D uses an ensemble of nine Long Short Term Memory (LSTM), Bidirectional Recurrent Neural Network (BRNN), and Residual Network (ResNet) models. SPOT-1D has 12,450 and 1250 proteins in its training and testing sets.

Recent inter-residue distance prediction algorithms include RaptorX [4], PDNET [31] and DeepDist [32]. PDNET and DeepDist both have MAE values 4.1Å whereas RaptorX [4] has MAE less than 4Å. So we chose RaptorX over others as it has less MAE. RaptorX predicts distances for residue pairs having at least 12 other residues in between in the sequence and within predicted distances less than or equal to 15Å. RaptorX uses an ensemble of three ResNet models and 2020 Cath S35 data for its training.

For residue-residue contact map prediction, recent methods include SPOT-Contact [17], RaptorX-Contact [33], Dncon2 [34]. Dncon2 obtains less than 70% precision. Raptor-X achieves less than 80% precision. SPOT-Contact gets more than 80% precision for top $L/10$ predictions for short, medium and long-range contacts. So we use SPOT-Contact for contact prediction. SPOT-Contact uses a coupling of residual 2D bidirectional LSTM with convolutional neural networks (CNN) and the same data set of SPOT-1D.

For secondary structure prediction, recent methods include PSIPred [35], DISTILL [36], and SSpro8 [37]. SSpro8 achieves the highest accuracy levels of about 92% and 79% respectively with and without using homologous proteins. So we choose SSpro8 for secondary structure prediction. SSpro8 uses an ensemble of BRNNs and 5772 training proteins.

B. USING OPTIMISATION SEARCH ALGORITHMS

We describe conformation representation, generation, and evaluation along with scoring functions used in our search.

We mainly use distance maps in scoring functions, but we also experiment with contact maps. We describe our scoring functions below where numeric parameters are fixed after preliminary experiments, but for the sake of brevity, we do not show those results. Note scores are not defined for sequentially proximate residues i and j with $i-j < 3$.

a: DISTANCE MAP BASED SCORING FUNCTION

For a pair of residues i and j, RaptorX [4] provides a predicted distance d_{ij} and a deviation δ_{ij} in the prediction. Using these, we define minimum and maximum allowable distances $m_{ij} = d_{ij} - \delta_{ij}$ and $M_{ij} = d_{ij} + \delta_{ij}$, and relative error $r_{ij} = \delta_{ij}/d(ij)$. Consequently, we do not include residue pairs for which relative errors are 0.5 or more. Next, for a current conformation c, we define a partial score s^c_{ij} and the total score s^c as below. FIGURE 4 (left) shows our distance map based scoring function s^c_{ij} with $m_{ij} = 5$ and $M_{ij} = 7$ for any residue pair i and j. As we see, the lower bound of the score for a residue pair is -1.

$$
\begin{aligned}
s^c_{ij} &= \frac{m_{ij} - d^c_{ij}}{m_{ij}} \quad when d^c_{ij} < m_{ij} \\
&= \frac{d^c_{ij} - M_{ij}}{M_{ij}} \quad when d^c_{ij} > M_{ij} \\
&= -1 \quad otherwise \\
s^c &= \sum_{ij} s^c_{ij} \quad where i-j \geq 3 \wedge r_{ij} < 0.5
\end{aligned}
$$

b: CONTACT MAP BASED SCORING FUNCTION

For a pair of residues i and j, SPOT-Contact [17] provides a predicted probability p_{ij} for the residue pair to be in contact.

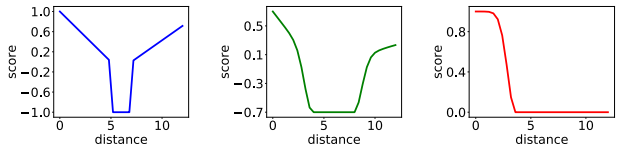


FIGURE 4. Scoring functions: distance map based (left), contact map based (middle), and steric clash based (right).

We consider residue pairs with contact probabilities at least 0.3. We give more emphasis on a greater probability. Moreover, we consider two residues are in contact when their distance is in between a minimum distance $d = 3.8$Å and a maximum distance $D = 8.0$Å. Next, for a current conformation c, we define a *partial score* σ_{ij}^c and the *total score* σ^c as below. FIGURE 4 (middle) shows our contact map based scoring function σ_{ij}^c with $d = 3.8$, $D = 8.0$, and $p_{ij} = 0.7$ for any residue pair i and j. As we see, the lower bound of the score for a pair of residues with indexes i and j is $-p_{ij}$. Note our scoring function is somewhat similar to the bounded potential [14] and the square well [38] functions.

$$\begin{aligned}\sigma_{ij}^c &= -p_{ij}e^{-(d-d_{ij}^c)^2} + p_{ij}\frac{d-d_{ij}^c}{d} \quad \text{when} d_{ij}^c < d\\ &= -p_{ij}e^{-(d_{ij}^c-D)^2} + p_{ij}\frac{d_{ij}^c-D}{d_{ij}^C} \quad \text{when} d_{ij}^c > D\\ &= -p_{ij} \quad otherwise\\ \sigma^c &= \sum_{ij}\sigma_{ij}^c \quad \text{where} i-j \geq 3 \wedge p_{ij} \geq 0.3\end{aligned}$$

c: STERIC CLASH BASED SCORING FUNCTION

Our distance or contact based scoring functions do not include all residue pairs. So to avoid steric clashes between residues, we define another scoring function. For this, we consider a clash when residue pairs have C^α atoms within $\Theta = 3.6$Å of each other. For a current conformation c, we define a partial score χ_{ij}^c and the total score χ^c as below. FIGURE 4 (right) shows our steric clash based scoring function χ_{ij}^c with $\Theta = 3.6$ for a given residue pairs with indexes i and j.

$$\begin{aligned}\chi_{ij}^c &= 1-e^{-(\Theta-d_{ij}^c)^2} \text{when } d_{ij}^c < \Theta\\ &= 0 \quad \text{otherwise}\\ \chi^c &= \sum_{ij}\chi_{ij}^c \quad \text{where} i-j \geq 3\end{aligned}$$

d: CONFORMATION REPRESENTATION

We primarily represent a conformation by the ϕ and ψ values of the residues. However, to use distance or contact maps, we also compute coordinates but only for N, C^α, C, and C^β atoms of each residue. During search when ϕ or ψ values are changed, to generate neighbour conformations, we recompute the coordinates of the atoms that will be affected by the changes.

e: OPTIMISATION SEARCH FRAMEWORK

Algorithm 1 shows the pseudocode of our simple local search algorithm for PSP. In this algorithm, geometric constraints learnt by machine learning algorithms have been turned into objective functions via the scoring functions. Local search algorithms usually generate neighbours randomly or in a generic way unrelated to the specific problem. Constraint guided sampling embeded within local search provides problem specific knowledge coded as constraints. Nevertheless, Algorithm 1 uses distance map based scoring function s^c but s^c could be easily replaced by contact map based scoring function σ^c. We further discuss the details of the algorithm.

The complexity of Algorithm 1 is $O(MNL^2)$ where M is the number of time iterations, N is the number of neighbouring conformation generated from each given conformation, and L is the number of AA residues in the given protein.

Algorithm 1 Constraint Guided Neighbours for PSP

1: Initialise conformation c from predicted ϕ, ψ values
2: Evaluate conformation c by computing scores s^c and χ^c
3: $\Delta\Phi \leftarrow \{\pm 3, \pm 6, \cdots, \pm\phi_{\textbf{MAE}}\}$ // SPOT-1D $\phi_{\textbf{MAE}} = 16$
4: $\Delta\Psi \leftarrow \{\pm 3, \pm 6, \cdots, \pm\psi_{\textbf{MAE}}\}$ // SPOT-1D $\psi_{\textbf{MAE}} = 23$
5: Reset tabu on each pair $\langle i, j\rangle$ appearing in s^c and χ^c
6: **for** iteration τ from 1 to a maximum M **do** // M = 8000
7: **MinClash** $\leftarrow (\chi^c > \chi_{\textbf{t}}) \wedge$ **probability**(0.5) // $\chi_{\textbf{t}} = 2$
8: **if MinClash then** // minimise clash based score
9: $\langle i, j\rangle \leftarrow$ **argmax**$_{\langle i',j'\rangle:\neg\textbf{tabu}(\langle i',j'\rangle,\tau)} \chi_{i'j'}^c$
10: **else** // minimise distance based score
11: $\langle i, j\rangle \leftarrow$ **argmax**$_{\langle i',j'\rangle:\neg\textbf{tabu}(\langle i',j'\rangle,\tau)} s_{i'j'}^c$
12: **end if**
13: $k \leftarrow$ **random**$(\{k' : i < k' < j \wedge$ **isLoop**$(k')\}$
14: $\Delta\phi \leftarrow N$ random values from $\Delta\Phi$ // N = 20
15: $\Delta\psi \leftarrow N$ random values from $\Delta\Psi$ // N = 20
16: $C \leftarrow \{c_n :$ add $\Delta\phi[n]$ to ϕ_k, $\Delta\psi[n]$ to ψ_k in $c\}$
17: Evaluate each $c_n \in C$ by computing scores s^{c_n}, χ^{c_n}
18: **if MinClash then** // minimise clash based score
19: $c' \leftarrow$ **argmin**$_{c_n} \chi^{c_n}$
20: Accept c' as c, **if** $\chi^{c'} < \chi^c \wedge s^{c'} \leq s^c$
21: **else** // minimise distance map based score
22: $c' \leftarrow$ **argmin**$_{c_n} s^{c_n}$
23: Accept c' as c, **if** $(s_{ij}^{c'} < s_{ij}^c \wedge \chi^{c'} < \chi_{\textbf{t}})$
24: **else** accept c' as c, **if** $(s^{c'} < s^c \wedge \chi^{c'} \leq \chi^c)$
25: **end if**
26: Apply tabu on $\langle i, j\rangle$ with tenure T // $T = 15$
27: **end for**
28: **return** the best 5 conformations in terms of s^c

f: CONFORMATION INITIALISATION

In Algorithm 1 Line 1, we take predicted ϕ and ψ values and MAE values ϕ_{MAE} and ψ_{MAE} from SPOT-1D [39]. We then generate random values from ranges $[\phi - \phi_{\text{MAE}}, \phi + \phi_{\text{MAE}}]$ and $[\psi - \psi_{\text{MAE}}, \psi + \psi_{\text{MAE}}]$. We also consider another alternative initialisation procedure: we use SSpro8 [37] predicted

TABLE 1. Typical ϕ and ψ ranges for secondary structures helixes (G, H, I), sheets (B, E), and loops (S, T, C) [40].

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

secondary structures and generate ϕ and ψ values randomly from the ranges shown in TABLE 1. Note once initialised, dihedral angles of only loop residues are changed by search. This is because SPOT-1D predictions for helix and sheet residues have smaller errors than those for loops.

g: CONFORMATION EVALUATION

In Algorithm 1 Lines 2 and 17, we evaluate a conformation c by computing the distance map based score s^c and the steric clash based score χ^c. We do not add s^c and χ^c, since their normalisation is not straightforward. Consequently, we have a two-objective minimisation problem, where s^c is the primary objective. However, at a time, we mainly work with one objective function, which is chosen in Line 7 in Algorithm 1. If χ^c is more than a threshold $\chi^{\mathsf{t}} = 2$, with 50% probability, we minimise χ^c; otherwise, we minimise s^c.

h: NEIGHBOUR GENERATION

Using the scoring function selected in Line 7 in Algorithm 1, we choose a residue pair i and j with the worst score (the tabu condition is discussed later) in Lines 9 and 11. In Line 13, we then choose a random loop residue k, which is in between i and j. Next, in Lines 14 and 15, we choose N angle differences in each of $\Delta\phi$ and $\Delta\psi$ respectively from sets $\Delta\Phi$ and $\Delta\Psi$. Note $\Delta\Phi$ and $\Delta\Psi$ as defined in Lines 3 and 4 hold values in intervals of 3° from ranges $[-\phi_{\text{MAE}}, +\phi_{\text{MAE}}]$ and $[-\psi_{\text{MAE}}, +\psi_{\text{MAE}}]$ respectively. Then, in Line 16, we generate N neighbour conformation using the angle differences in $\Delta\phi$ and $\Delta\psi$.

i: USING TABU METAHEURISTIC

Revisitation is a problematic issue in local search. In Algorithm 1 Lines 9 and 11, the same i and j could be repeatedly selected. To avoid revisitation, we use the tabu metaheuristic [41]. With tabu initialised in Line 5, enforced in Line 26, and checked via $\mathsf{tabu}(\langle i', j' \rangle, \tau)$ in Lines 9 and 11, recently selected i and j will not be selected again in Lines 9 and 11 within a number (called tabu tenure T) of future iterations. In this work, we do not apply tabu on the selection of residue k in Line 13.

j: ACCEPTING BEST NEIGHBOUR

When we improve one objective function, we do not want to worsen the other one. Moreover, improving the partial score $s_{ij}^{c'}$ is the primary reason to select the residues i and j in Line 11. So when distance map based scoring function s^c is chosen in Line 7, we accept neighbour c' with best $s^{c'}$ in Line 24, if the partial score $s_{ij}^{c'}$ is strictly better than s_{ij}^{c}. We also alternatively accept c' when $s^{c'}$ is strictly better than s^c and $\chi^{c'}$ is not worse than χ^c. Next, steric clash minimisation is basically a secondary objective. So when steric clash based score χ^c is chosen in Line 7, we accept neighbour c' with best $\chi^{c'}$ in Line 20, if $\chi^{c'}$ is strictly better than χ^c and $s^{c'}$ is not worse than s^c.

k: IMPLEMENTATION PLATFORM

We implement our algorithms on top of a recently developed Python-based PSP search platform named Koala, which draws concepts from a constraint based local search system named Kangaroo [42].

VI. EXPERIMENTAL RESULTS

All the algorithms are executed on a Linux 64-bit system with Intel® Xeon® X3470 293 X 8 GHz and 8GB memory.

TABLE 2 shows our experimental results on 35 proteins having 42 to 138 residues.

A. DATASET

Our dataset includes 14 α type, 11 β type, and 10 α/β type proteins. These proteins are from existing PSP search algorithms such as QUARK [5], MODE-K [9], and MODCSA/CA [43] or a machine learning algorithm such as SPOT-1D [39]. We have used CD-HIT and PSI-BLAST [44] to ensure the proteins do not have more than 25% sequence similarity with the training proteins of the previously-mentioned machine learning algorithms used in our implementation.

B. COMPARISON OF OUR ALGORITHM VERSIONS

Besides the steric clash based scoring function χ^c, Algorithm 1 uses (*i*) distance map based scoring function s^c, (*ii*) tabu with tenure 15, (*iii*) initialisation using predictions from SPOT-1D [39], (*iv*) selection of residue pairs i and j based on scoring functions χ_{ij}^c or s_{ij}^c, and (v) generation of $\Delta\phi$ and $\Delta\psi$ values from the ranges determined by the MAE values of SPOT-1D. To test the effectiveness of each of the components mentioned, we create the following 7 versions of the proposed algorithm.

- **dm**: is the exact algorithm as is described in Algorithm 1 with the 5 components mentioned above.
- **cm**: uses the contact map based scoring function σ^c instead of the distance map based scoring function s^c in **dm**.
- **nt**: does not use the tabu metaheuristic used in **dm** and so more revisitation of selected residue pairs could occur.
- **rp**: selects residue pairs i and j randomly but still satisfying the condition $i-j \geq 3 \wedge r_{ij} < 0.5$ as is needed in the definition of the distance map based scoring function.
- **rl**: randomly selects a loop region first and then a random residue k from that loop. Note **dm** first selects residue pairs i and j using chosen scoring functions

TABLE 2. Top: mean RMSD values obtained for proteins (left) by proposed algorithm variants (center) and state-of-the-art algorithms (right). Bottom: the number of proteins with mean RMSD values the best (emboldened) and the 2nd best (underlined), and also the number of proteins with mean RMSD values ≤ various threshold levels when our algorithm variants are compared with each other and when our best version is compared with the state-of-the-art algorithms. Note that CGNP is actually dm.

Type	Protein	Length	cm	nt	rp	rl	ri	fr	dm	CGNP	CGLFOLD	trRosettaX
α	5AON	48	3.560	4.432	5.167	4.447	**2.224**	4.490	3.246	**3.246**	6.412	4.332
	5B1A	58	**7.905**	10.918	9.693	10.328	10.745	9.605	9.547	9.547	17.140	**5.316**
	1SXD	91	8.905	9.414	**3.542**	7.198	9.693	6.118	5.545	**5.545**	9.060	11.553
helixes	5B1N	59	3.973	4.230	4.409	4.891	4.406	3.790	**3.622**	**3.622**	4.430	3.718
and	5COS	56	3.965	4.615	4.425	4.700	4.042	4.265	**3.000**	**3.000**	3.130	4.414
loops	5E5Y	61	9.317	6.087	9.300	10.109	6.654	5.751	**5.354**	**5.354**	6.025	8.920
	5FVK	82	4.667	6.947	4.627	8.134	8.929	3.117	**3.099**	**3.099**	3.570	4.596
	5EMX	54	6.099	5.202	4.517	5.917	6.515	4.430	**4.126**	**4.126**	5.544	4.446
	5TDY	42	8.096	6.796	8.404	7.560	6.893	7.444	**6.650**	**6.650**	10.000	7.129
	5HE9	56	5.687	6.008	5.762	6.028	7.151	**4.991**	5.834	**5.834**	8.252	6.970
	2O4T	90	7.938	7.612	8.764	11.400	9.450	7.324	**6.557**	**6.557**	10.682	9.929
	2O42	138	15.262	15.340	**11.287**	14.545	14.392	12.143	12.785	12.785	13.624	**9.812**
	5B5I	67	10.369	8.134	8.899	11.192	8.302	**6.008**	8.407	8.407	9.860	**7.088**
	5DIC	115	8.591	13.014	8.164	9.223	13.126	7.995	**4.792**	4.792	**3.332**	8.49
β	1R75	110	9.101	9.903	10.300	10.024	13.404	10.024	**7.238**	7.238	13.078	**6.518**
	1OK0	74	7.075	7.700	7.220	7.882	6.280	8.326	**5.308**	**5.308**	7.851	11.876
	2AXW	134	13.712	13.937	**11.610**	15.677	13.673	12.201	12.719	12.719	15.466	**8.486**
sheets	2BT9	90	8.601	6.770	8.300	9.398	**5.228**	7.422	6.317	**6.317**	6.574	7.028
and	2CHH	113	18.404	18.602	**16.550**	17.786	24.055	18.934	18.860	18.86	**8.565**	11.348
loops	2V33	91	11.512	9.368	10.000	12.286	9.965	9.464	**7.486**	7.486	**7.381**	9.488
	5AEJ	113	**16.989**	19.667	21.159	20.000	20.037	24.432	22.261	22.261	17.074	**7.496**
	5AOT	102	19.501	**8.991**	12.100	13.999	14.403	11.621	9.387	**9.387**	12.233	9.642
	5EZU	67	7.679	7.302	8.610	8.681	7.213	7.517	**5.848**	**5.848**	7.526	6.921
	5FUI	124	13.835	15.054	12.000	15.420	19.214	13.539	**11.730**	11.730	**11.381**	11.974
	5HDW	131	12.274	16.737	12.199	13.426	13.903	10.159	**10.058**	**10.058**	12.095	10.310
α/β	1CRN	46	6.311	4.470	4.913	6.895	**3.331**	5.800	4.260	**4.260**	4.8372	4.602
	1CF7	82	7.998	7.245	7.452	8.658	7.056	**4.957**	6.762	6.762	**4.600**	8.938
	1IS7	84	8.284	7.656	9.460	8.411	8.837	9.512	**4.644**	**4.644**	7.499	6.603
helixes	1KA8	100	11.582	10.259	11.780	11.674	9.992	9.820	**6.650**	**6.650**	8.722	9.604
sheets	1MC2	122	10.241	14.405	**8.642**	10.656	9.611	10.322	8.760	8.760	10.29	**5.065**
looops	1T1J	125	7.808	11.320	7.830	12.776	10.427	8.493	**5.702**	**5.702**	6.471	10.612
	1Y71	112	12.598	13.532	10.508	12.234	12.245	10.255	**7.592**	**7.592**	7.778	9.962
	2BSE	107	11.578	15.128	10.074	13.012	**9.661**	10.699	10.237	10.237	10.257	**7.480**
	3BJO	100	**8.442**	11.080	10.100	12.811	10.509	9.621	10.062	10.062	9.016	**8.921**
	3CHB	103	12.887	14.361	**11.126**	12.932	13.565	12.375	11.589	11.589	**8.961**	9.334
Best RMSD			3	1	6	0	4	3	**18**	**20**	6	9
2nd Best RMSD			5	5	3	0	2	**11**	9	**11**	11	**13**
RMSD ≤ 6Å			5	5	8	4	5	9	**14**	**14**	7	8
RMSD ≤ 9Å			19	17	18	13	15	18	**24**	**24**	20	22
RMSD ≤ 12Å			26	24	**31**	22	24	**29**	**31**	31	28	**35**

and then selects a loop residue in between residues i and j.

ri: unlike **dm**, initialises the ϕ and ψ angles randomly but using the SS specific angle ranges shown in TABLE 1.

fr: like **ri**, initialises the ϕ and ψ angles randomly from the SS specific angle ranges and unlike **dm**, generates $\Delta\phi$ and $\Delta\psi$ values from the full range of $[-180°, +180°]$.

We run each of the 7 versions of our algorithm on each protein 5 times. Each run has the maximum iteration $M = 8000$ and the number of neighbours generated in each iteration $N = 20$. So each run essentially explores 160, 000 conformations; this is the same number of conformations explored by CGLFOLD [3]. Nevertheless, from each run, we take 5 best conformations in terms of the respective distance or contact map based scoring function used. Then, we compute mean Root Mean Square Deviation (RMSD) value over the 25 conformations for each protein for the same algorithm version and show in TABLE 2 (top left).

Among our 7 versions, as we see in TABLE 2 (bottom left), **dm** obtains the best mean RMSD values in 18 out of 35 proteins and 2nd best mean RMSD values in 9 proteins. We perform Wilcoxon signed rank test with 95% confidence interval on **dm** against the other 6 versions and p-values are at most 0.0008. This indicates **dm**'s performance is statistically significantly different from the other versions. Moreover, TABLE 2 (bottom) also shows the numbers of proteins in which various versions obtain mean RMSD values ≤ various threshold values such as 6Å, 9Å, and 12Å. Clearly, **dm** obtains the best performance among the versions particularly with thresholds 6Å and 9Å. From these results, it is clear that each component of **dm** is important for its performance. We will perform further analysis later in the paper.

Henceforth, we name our best algorithm version **dm** as Constraint Guided Neighbours for PSP (CGNP).

C. COMPARISON WITH STATE-OF-THE-ART METHODS

We compare our proposed CGNP with two most related recent PSP search methods CGLFOLD [3] and trRosettaX [12]. CGLFOLD performs perturbation based loop sampling along with predicted contact map based scoring function. On the other hand, trRosettaX performs gradient minimisation along with a scoring function that has components based on predicted distance maps and inter-residue angle orientations. Of course both CGLFOLD and trRosettaX randomly generate neighbour conformations.

For running CGLFOLD on the proteins and computing mean RMSD values, we use the same setting that we have used in the experiments with our algorithm versions. For trRosettaX, we use only the distance based component in the scoring function, since CGNP uses distance based scores. Note that our main objective in this paper is not to explore scoring functions but is rather to see the effectiveness of our constraint-guided neighbour generation approach over existing random-based approaches. However, to investigate that, we do need a scoring function and we use distant based ones. Nevertheless, trRosettaX returns just one conformation per run. So we run trRosettaX on each protein 25 times and compute mean RMSD values over the 25 conformations.

TABLE 2 (top right) shows the mean RMSD values obtained by CGNP, CGLFOLD, and trRosettaX in each protein. As we see, CGNP obtains best mean RMSD values in more proteins in all three types α, β, and α/β proteins than CGLFOLD and trRosettaX do. Overall, TABLE 2 (bottom right) shows CGNP obtains best mean RMSD values in 20 out of 35 proteins and 2nd best RMSD values in 11 proteins. We perform Friedman test with 95% confidence level on CGNP, CGLFOLD, and trRosettaX performances and get p-value 0.0027. Then, for posthoc analysis, we perform Nemenyi test with 95% confidence level to compute pairwise differences among the three algorithms. From the test results, we see that CGLFOLD and trRosettaX have no statistically significantly difference with p-value 0.6046 but CGNP is statistically significantly different from CGLFOLD and trRosettaX with p-values 0.0026 and 0.0444 respectively.

We run all the algorithms on three proteins of three different types and check their running time in table TABLE 3. Note that these algorithms have been implemented on different platforms and programming languages. For example, our method and CGLFOLD are implemented on Python, which as a programming language and platform is by default slow. On the other hand trRosettaX is implemented on C/C++ programming language and is so inherently fast.

FIGURE 5 shows the best conformations obtained by CGNP, CGLFOLD, and trRosettaX for a sample protein 1IS7.

D. FURTHER PERFORMANCE ANALYSIS

FIGURE 6 shows the correlation between the distance map based scores and the RMSD values of the conformations generated during search in the sample runs of CGNP on a sample protein 1IS7 of type α/β. The Pearson correlation coefficient between the scores and the RMSD values is 0.665. These results show that improving the distance map based scores could highly likely lead to improving better conformations in terms of RMSD values.

FIGURE 7 shows the differences in mean RMSD values of the initial and the final conformations for CGNP. We see that CGNP statistically significantly (Wilcoxon signed rank test 95% confidence level p-value 0.0000) improves the quality of the conformations in terms of the RMSD values.

FIGURE 8 shows the best distance map based scores obtained so far in each iteration of sample runs of rp, rl, and dm versions of our algorithm for a sample protein 1IS7. Clearly, dm keeps improving the distance map based scores while rp and rl get somewhat stuck in plateaus in terms of achieving better scores. These results show the effectiveness

TABLE 3. Running time analysis.

Protein	Type	Length	Method	Time
1CRN	α	46	CGNP	23 mins 39s
			CGLFOLD	1 hr 41 mins
			trRosettaX	3 mins 46s
5AEJ	β	113	CGNP	3 hrs 20 mins
			CGLFOLD	4 hrs 70min
			trRosettaX	30 mins 20s
2O42	$\alpha\ \beta$	138	CGNP	2 hrs 10 mins
			CGLFOLD	3 hrs 20 mins
			trRosettaX	25 mins

FIGURE 5. Best conformations by CGNP, CGLFOLD, and trRosettaX (all cyan) w.r.t. native conformations (green) for protein 1IS7 of Type α/β and length 84.

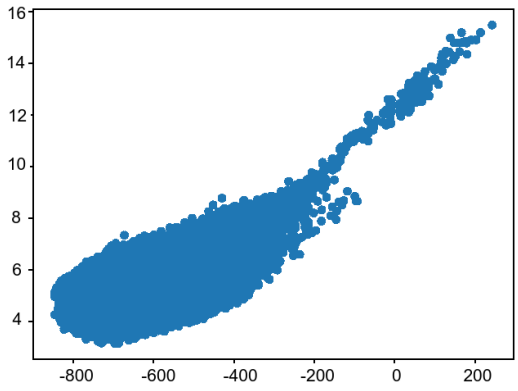


FIGURE 6. Scatter plots of distance map based scores (x-axis) vs RMSD values (y-axis) for α/β type protein 1IS7.

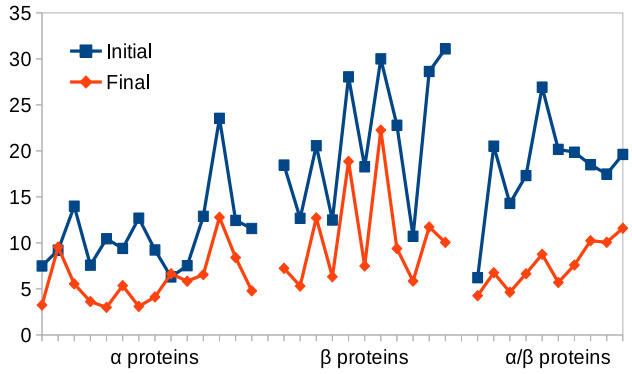


FIGURE 7. Differences in mean RMSD values of the initial and the final conformations returned by CGNP.

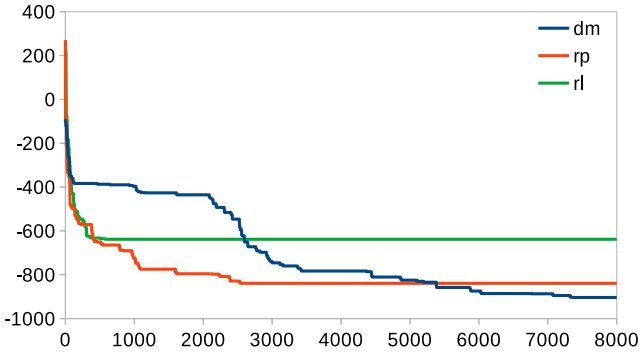


FIGURE 8. Best distance map based scores obtained so far (y axis) in each iteration of sample runs of rp, rl, and dm versions of our algorithm for α/β type protein 1IS7.

of our constraint-guided neighbour generation approach of dm over the random selection based approaches of rp and rl in terms of improving the distance map based scores.

FIGURE 9 shows sample RMSD distributions of the conformations generated by the sample runs of the rp, rl, and dm versions of our algorithm for one sample protein 1IS7. These three versions are for the various ways we select the residue pairs or the loop regions to eventually select another residue of which the ϕ and ψ angles will be changed. Clearly, selecting a random loop region by rl is the worst among the three versions as rl explores inferior conformations. Between selecting a random pair by rp and a greedy pair by dm, the greedy pair selection explore more promising conformations in most cases. These results show the effectiveness of our constraint-based conformation generation approach in dm over the random selection based approaches in rp and rl in terms of exploring higher quality conformations.

VII. CASP13 AND CAMEO144 PROTEINS

We have also run our method with the same experimental setting as describe before on 20 proteins from CASP13 protein and CAMEO144 hard target test set and compared it with a very recent method trRosettaX [12] and reported RMSD and GDT-TS score. GDT-TS score has been used in ranking PSP

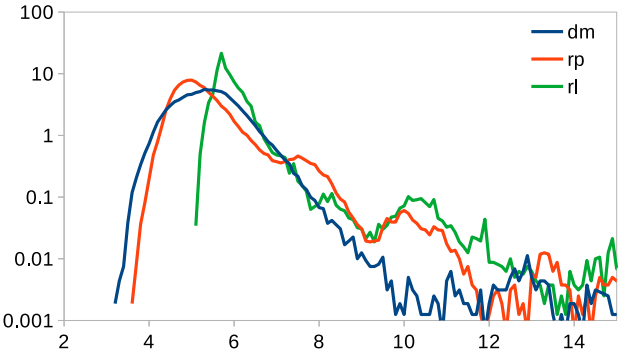


FIGURE 9. Sample distribution of RMSD values (x axis) versus percentage of conformations generated (y-axis) by rp, rl, and dm versions of our algorithm for α/β type protein 1IS7.

TABLE 4. Mean RMSD and GDT-TS score values obtained for proteins by proposed algorithm and state-of-the-art algorithm.

Protein Type	Protein ID	Protein Length	RMSD		GDT	
			CGNP	trRosettaX	CGNP	trRosettaX
α	6SSR	136	**4.347**	7.086	**0.718**	0.369
	7JRQ	127	**5.23**	5.807	0.56	**0.554**
	6VAX	113	**7.865**	9.148	**0.498**	0.419
helixes	6T3I	86	**3.843**	5.103	**0.61**	0.58
and loops	6SIF	51	**3.073**	5.946	**0.686**	0.571
	6V8H	141	**7.3**	8.632	**0.59**	0.547
	T0957S2	157	**8.921**	10.322	**0.311**	0.199
β	7C28	65	**4.185**	6.512	0.577	**0.527**
	7CCB	164	9.778	**9.355**	0.454	**0.463**
sheets	T0968S2	114	**8.316**	13.028	**0.388**	0.319
and loops	T0992	107	**4.469**	6.679	**0.56**	0.54
	T0981	203	**10.23**	10.595	0.34	**0.362**
α/β	6I1M	93	**7.1**	**0.493**	4.8372	0.477
	6L7Q	145	**12.919**	12.986	0.289	0.289
	6LXG	90	**3.522**	4.201	**0.68**	0.67
helixes	6UXC	172	**10.488**	10.503	0.352	**0.516**
sheets	T0949	183	**15.865**	19.655	**0.183**	0.178
and loops	T0958	84	**5.669**	6.393	**0.5**	0.39
	T0997	185	**8.91**	8.926	**0.45**	0.446
	T1022S1	223	**14.342**	14.668	**0.288**	0.277

methods that took part in CASP14. In most of the proteins as shown in TABLE 4, our method achieve better result than trRosettaX.

VIII. CONCLUSION

Protein structure prediction (PSP) has achieved significant progress lately via development of geometric constraint based scoring functions. However, sample generation for PSP remains challenging as existing search algorithms take random based approaches. We propose a constraint-guided novel approach to identify problematic parts of a current conformation and then to make changes to those parts to generate neighbour conformations. Our approach thus makes informed decisions in neighbour generation and explains its performance. On a set of benchmark proteins of varying types and sizes, our approach significantly outperforms state-of-the-art PSP search algorithms that use random sampling with similar scoring functions.

ACKNOWLEDGMENT

(Rianon Zaman and M. A. Hakim Newton are co-first authors.)

REFERENCES

[1] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, Jul. 2021.

[2] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, "Improved protein structure prediction using predicted interresidue orientations," *Proc. Nat. Acad. Sci. USA*, vol. 117, no. 3, pp. 1496–1503, Jan. 2020.

[3] J. Liu, X.-G. Zhou, Y. Zhang, and G.-J. Zhang, "CGLFold: A contact-assisted de novo protein structure prediction using global exploration and loop perturbation sampling algorithm," *Bioinformatics*, vol. 36, no. 8, pp. 2443–2450, Apr. 2020.

[4] J. Xu, "Distance-based protein folding powered by deep learning," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 34, pp. 16856–16865, Aug. 2019.

[5] D. Xu and Y. Zhang, "*Ab initio* protein structure assembly using continuous structure fragments and optimized knowledge-based force field," *Proteins, Struct., Function, Bioinf.*, vol. 80, no. 7, pp. 1715–1735, Jul. 2012.

[6] X.-G. Zhou, C.-X. Peng, J. Liu, Y. Zhang, and G.-J. Zhang, "Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction," *IEEE Trans. Evol. Comput.*, vol. 24, no. 3, pp. 536–550, Jun. 2020.

[7] S. Song, S. Gao, X. Chen, D. Jia, X. Qian, and Y. Todo, "AIMOES: Archive information assisted multi-objective evolutionary strategy for ab initio protein structure prediction," *Knowl.-Based Syst.*, vol. 146, pp. 58–72, Apr. 2018.

[8] S. H. P. de Oliveira, E. C. Law, J. Shi, and C. M. Deane, "Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction," *Bioinformatics*, vol. 34, no. 7, pp. 1132–1140, Apr. 2018.

[9] X. Chen, S. Song, J. Ji, Z. Tang, and Y. Todo, "Incorporating a multi-objective knowledge-based energy function into differential evolution for protein structure prediction," *Inf. Sci.*, vol. 540, pp. 69–88, Nov. 2020.

[10] L. D. L. Corrêa, B. Borguesan, M. J. Krause, and M. Dorn, "Three-dimensional protein structure prediction based on memetic algorithms," *Comput. Oper. Res.*, vol. 91, pp. 160–177, Mar. 2018.

[11] L. de Lima Corrêa and M. Dorn, "A multi-population memetic algorithm for the 3-D protein structure prediction problem," *Swarm Evol. Comput.*, vol. 55, Jun. 2020, Art. no. 100677.

[12] H. Su, W. Wang, Z. Du, Z. Peng, S. Gao, M. Cheng, and J. Yang, "Improved protein structure prediction using a new multi-scale network and homologous templates," *Adv. Sci.*, vol. 8, no. 24, Dec. 2021, Art. no. 2102592.

[13] B. R. Brooks *et al.*, "CHARMM: The biomolecular simulation program," *J. Comput. Chem.*, vol. 30, no. 10, pp. 1545–1614, 2009.

[14] A. Leaver-Fay *et al.*, "ROSETTA3: An object-oriented software suite for the simulation and design of macromolecules," in *Methods in Enzymology*, vol. 487. New York, NY, USA: Academic, 2011, pp. 545–574.

[15] J. Zhang and Y. Zhang, "A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction," *PLoS ONE*, vol. 5, no. 10, Oct. 2010, Art. no. e15386.

[16] T. J. Richmond, "Solvent accessible surface area and excluded volume in proteins: Analytical equations for overlapping spheres and implications for the hydrophobic effect," *J. Mol. Biol.*, vol. 178, no. 1, pp. 63–89, 1984.

[17] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou, "Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks," *Bioinformatics*, vol. 34, no. 23, pp. 4039–4045, Dec. 2018.

[18] B. Adhikari, D. Bhattacharya, R. Cao, and J. Cheng, "CONFOLD: Residue-residue contact-guided *ab initio* protein folding," *Proteins, Struct., Function, Bioinf.*, vol. 83, no. 8, pp. 1436–1449, 2015.

[19] B. Adhikari and J. Cheng, "CONFOLD2: Improved contact-driven *ab initio* protein structure modeling," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–5, 2018.

[20] J. Hou, T. Wu, R. Cao, and J. Cheng, "Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13," *Proteins, Struct., Function, Bioinf.*, vol. 87, no. 12, pp. 1165–1178, Dec. 2019.

[21] L. A. Abriata, G. E. Tamò, B. Monastyrskyy, A. Kryshtafovych, and M. D. Peraro, "Assessment of hard target modeling in CASP12 reveals an emerging role of alignment-based contact prediction methods," *Proteins, Struct., Function, Bioinf.*, vol. 86, pp. 97–112, Mar. 2018.

[22] J. Schaarschmidt, B. Monastyrskyy, A. Kryshtafovych, and A. M. J. J. Bonvin, "Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age," *Proteins, Struct., Function, Bioinf.*, vol. 86, pp. 51–66, Mar. 2018.

[23] J. Ma, S. Wang, Z. Wang, and J. Xu, "Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning," *Bioinformatics*, vol. 31, no. 21, pp. 3506–3513, Nov. 2015.

[24] S. Wang, W. Li, R. Zhang, S. Liu, and J. Xu, "CoinFold: A web server for protein contact prediction and contact-assisted protein folding," *Nucleic Acids Res.*, vol. 44, no. W1, pp. W361–W366, Jul. 2016.

[25] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLOS Comput. Biol.*, vol. 13, no. 1, Jan. 2017, Art. no. e1005324.

[26] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020.

[27] V. Cutello, G. Narzisi, and G. Nicosia, "A multi-objective evolutionary approach to the protein structure prediction problem," *J. Roy. Soc. Interface*, vol. 3, no. 6, pp. 139–151, Feb. 2006.

[28] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "Hierarchical structure of proteins," in *Molecular Cell Biology*, 4th ed. San Francisco, CA, USA: Freeman, 2000.

[29] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, and Y. Zhou, "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* (Methods in Molecular Biology), vol. 1484. Totowa, NJ, USA: Human Press, 2017, pp. 55–63.

[30] F. Mataeimoghadam, M. A. H. Newton, A. Dehzangi, A. Karim, B. Jayaram, S. Ranganathan, and A. Sattar, "Enhancing protein backbone angle prediction by using simpler models of deep neural networks," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Dec. 2020.

[31] B. Adhikari, "A fully open-source framework for deep learning protein real-valued distances," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 13374.

[32] B. Adhikari, "REALDIST: Real-valued protein distance prediction," *bioRxiv*, 2020, doi: 10.1101/2020.11.28.402214.

[33] S. Wang, S. Sun, Z. Li, R. Zhang, and J. Xu, "Accurate de novo prediction of protein contact map by ultra-deep learning model," *PLOS Comput. Biol.*, vol. 13, no. 1, Jan. 2017, Art. no. e1005324.

[34] B. Adhikari, J. Hou, and J. Cheng, "DNCON2: Improved protein contact prediction using two-level deep convolutional neural networks," *Bioinformatics*, vol. 34, no. 9, pp. 1466–1472, May 2018.

[35] D. W. A. Buchan, F. Minneci, T. C. O. Nugent, K. Bryson, and D. T. Jones, "Scalable web services for the PSIPRED protein analysis workbench," *Nucleic Acids Res.*, vol. 41, no. W1, pp. W349–W357, Jul. 2013.

[36] C. Mirabello and G. Pollastri, "Porter, PaleAle 4.0: High-accuracy prediction of protein secondary structure and relative solvent accessibility," *Bioinformatics*, vol. 29, no. 16, pp. 2056–2058, Aug. 2013.

[37] C. N. Magnan and P. Baldi, "SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity," *Bioinformatics*, vol. 30, no. 18, pp. 2592–2597, 2014.

[38] T. Kosciolek and D. T. Jones, "De novo structure prediction of globular proteins aided by sequence variation-derived contacts," *PLoS ONE*, vol. 9, no. 3, Mar. 2014, Art. no. e92197.

[39] J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou, "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*, vol. 35, no. 14, pp. 2403–2410, Jul. 2019.

[40] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolym., Original Res. Biomol.*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.

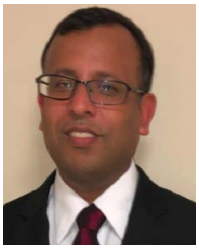
[41] F. Glover, "Tabu search methods in artificial intelligence and operations research," *ORSA Artif. Intell.*, vol. 1, no. 2, p. 6, 1987.

[42] M. H. Newton, D. N. Pham, A. Sattar, and M. Maher, "Kangaroo: An efficient constraint-based local search system using lazy propagation," in *Proc. Int. Conf. Princ. Pract. Constraint Program.* New York, NY, USA: Springer, 2011, pp. 645–659.

[43] D. Ramyachitra and A. Ajeeth, "MODCSA-CA: A multi objective diversity controlled self adaptive cuckoo algorithm for protein structure prediction," *Gene Rep.*, vol. 8, pp. 100–106, Sep. 2017.

[44] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–3402, 1997.

RIANON ZAMAN received the B.Sc. degree from the Khulna University of Technology (KUET) and the M.Sc. degree from United International University, Bangladesh. She is currently pursuing the Ph.D. degree with the School of ICT at Griffith University, Australia. Her research interests include bioinformatics, optimization, and machine learning.



M. A. HAKIM NEWTON received the B.Sc.Engg. and M.Sc.Engg. degrees from the Bangladesh University of Engineering and Technology (BUET) and the Ph.D. degree from Strathclyde University, U.K. He was a Research Engineer at the National ICT Australia (NICTA). He is a Lecturer with the School of Information and Physical Sciences, The University of Newcastle, Australia and also an Adjunct Senior Research Fellow with the Institute for Integrated and Intelligent Systems (IIIS), Griffith University, Australia. His research interests include artificial intelligence, intelligent search, machine learning, and bioinformatics.

FERESHTEH MATAEIMOGHADAM received the bachelor's degree from the Ferdowsi University of Mashhad, Iran, and the master's degree from Kharazmi University, Iran. She is currently pursuing the Ph.D. degree with the School of ICT, Griffith University, Australia. Her research interests include data science pipelines for biological data, optimization, statistical learning, and mathematical modeling for regression, classification, and clustering.

ABDUL SATTAR is a Professor with the School of ICT, Griffith University, Australia, where he was the founding Director of the Institute for Integrated and Intelligent Systems. He was also the Education Director at the Queensland Research Laboratory (QRL), National ICT Australia (NICTA). He won a number of ARC discovery grants and international awards for his work in artificial intelligence.