Constraint Guided Neighbour Generation for Protein Structure Prediction

RIANON ZAMAN1,*, M A HAKIM NEWTON2,*, FERESHTEH MATAEIMOZHADAM1, ABDUL SATTAR1,2
1School of Information and Communication Technology, Griffith University, Australia
2Institute for Integrated and Intelligent Systems, Griffith University, Australia
*These two authors have made equal contributions to this paper.

Corresponding author: Rianon.Zaman (e-mail: rianon.zaman@griffithuni.edu.au) and M A Hakim Newton (email: mahakim.newton@griffith.edu.au)

This research is partially supported by Australian Research Council Discovery 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 conformational search methods 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. 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.

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 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. So 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]–
Nevertheless, energy functions that involve all-atomic details used in PSP research is the ROSETTA [14] energy function. CHARMM [13]. Another such successful scoring function energy components have been used in protein structure chemical (bond energies), and electrostatic (Coulomb forces) known or defined so far, but physical (Van der Walls forces), the search and optimisation approach for PSP. Considering the relevance with this work, we mainly explore idea/implementation, experimental results, and conclusions.

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 details the problem formulation, the idea/implementation, experimental results, and conclusions.

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

The kinetic 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 considered very effective in PSP problem. 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.

Sequential search is used in SAINT2 [8], in which an independent fragment-assembly structure predictor predicts 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 that the angle Probability List strategy is quite useful in order to identify distinct structural patterns. As a energy function, it uses ROSETTA [14] and solvent accessible surface area (SASA) [16].

tRosetta [2], [12] claims that gradient descent algorithm is useful in solving the PSP problem. As energy function, tRosetta uses the 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 8Å). In distance and contact maps, residues are represented by Cβ atoms except by Cα 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] RaptorX [23]–[25] and AlphaFold [26]
predict distance maps and use them in their search algorithms for protein structures.

III. PROBLEM FORMULATION

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

FIGURE 1 Left shows AAs comprise $N$, $C^\alpha$, $C$, $O$, and $C^\beta$ atoms among others. $C^\alpha$ 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 $C^\alpha$ atom and $C^\beta$ 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 $\phi$, $\psi$, and $\omega$ that allow folding. These three angles are respectively defined by each four successive atoms from the sequence $C_{i-1}$, $N_i$, $C^\alpha_i$, $C_i$, $N_{i+1}$, $C^\alpha_{i+1}$. For most proteins, $\omega$ is $180^\circ$ [27], but $\phi$ and $\psi$ can take any values from $-180^\circ$ to $+180^\circ$. The side chains at individual AAs have their own dihedral angles, but in this work, we mainly focus on searching for backbone $\phi$ and $\psi$ angles of the main chain. Using backbone $\phi$ and $\psi$ 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 $\phi$ and $\psi$ values from very narrow ranges of about 20$^\circ$. Finding the $\phi$ and $\psi$ values for the flexible loop regions is challenging since they can take any fractional values in $[-180^\circ, +180^\circ]$. 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 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 $C^\alpha$ atoms except by $C^\alpha$ 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 helices 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 $d'_{ij}$ 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 $d'_{ij} = 9.25$ 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 $d'_{ij}$ 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.

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. 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 $\phi$ and $\psi$ 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.

Many inter-residue distance prediction algorithms have been proposed in recent few years. These 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 them in the sequence and within predicted distances less than or equal to 15Å.

For residue-residue contact map prediction, there are several methods, e.g. SPOT-Contact [17], RaptorX-Contact [33]. Dncon2 [34]. Dncon2 obtains less than 70% precision. Raptor-X achieves less than 80% precision, and SPOT-Contact gets more than 80% precision for top L/10 predictions for short, medium and long-range contacts, where L denotes the number of residues in the protein. So we use SPOT-Contact for contact prediction.

For secondary structure prediction, a few methods have been proposed: PSIPred [35], DISTILL [36]. SSpro8 [37] 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.

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 $\delta_{ij}$ 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 $s_{ij}^c$ and the total score $s^c$ as below. FIGURE 4 (left) shows our distance map based scoring function $s_{ij}^c$ 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 $-1$.

$$s_{ij}^c = \frac{m_{ij} - d_{ij}}{m_{ij}} \text{ when } d_{ij} < m_{ij}$$
$$s_{ij}^c = \frac{d_{ij} - M_{ij}}{M_{ij}} \text{ when } d_{ij} > M_{ij}$$
$$s^c = \sum_{ij} s_{ij}^c \text{ otherwise where } i - j \geq 3 \land r_{ij} < 0.5$$

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. 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 $d = 3.8\text{Å}$ and a maximum distance $D = 8.0\text{Å}$. Next, for a current conformation $c$, we define a partial score $\sigma_{ij}^c$ and the total score $s^c$ as below. FIGURE 4 (middle) shows our contact map based scoring function $\sigma_{ij}^c$ 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.

FIGURE 3. Our PSP pipeline

FIGURE 4. Scoring functions: distance map based (left), contact map based (middle), and steric clash based (right)
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 $C^\alpha$ atoms within $\Theta = 3.6\text{Å}$ of each other. For a current conformation $c$, we define a partial score $\chi_i^c$ and the total score $\chi^c$ as below.

\[
\chi_i^c = 1 - e^{-(\Theta - d_i^c)^2} \quad \text{when} \quad d_i^c < \Theta \\
0 \quad \text{otherwise}
\]

\[
\chi^c = \sum_i \chi_i^c \quad \text{where} \quad i-j \geq 3 \land p_{ij} \geq 0.3
\]

c: Steric Clash Based Scoring Function.

d: Conformation Representation.

We primarily represent a conformation by the $\phi$ and $\psi$ values of the residues. However, to use distance or contact maps, we also compute coordinates but only for $N$, $C^\alpha$, $C$, and $C^\beta$ atoms of each residue. During search when $\phi$ or $\psi$ 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 embedded within local search provides problem specific knowledge 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 $\sigma^c$. We further discuss the details of the algorithm.

f: Conformation Initialisation.

In Algorithm 1 Line 1, we take predicted $\phi$ and $\psi$ values and MAE values $\phi_{\text{MAE}}$ and $\psi_{\text{MAE}}$ 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 secondary structures and generate $\phi$ and $\psi$ 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 $\chi^c$. We do not add $s^c$ and $\chi^c$, since their normalisation is not straightforward. Consequently, we have 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 $\chi^c$ is more than a threshold $\chi = 2$, with 50% probability, we minimise $\chi^c$; 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

\[
\sigma_i^c = p_{ij} e^{-(d_i^c - d_{ij}^c)^2} + p_{ij} \frac{d - d_i^c}{d - d_{ij}^c} \quad \text{when} \quad d_i^c < d \\
- p_{ij} e^{-(d_i^c - d_{ij}^c)^2} + p_{ij} \frac{d - d_i^c}{d - d_{ij}^c} \quad \text{when} \quad d_i^c > D \\
- p_{ij} \quad \text{otherwise}
\]

\[
\sigma^c = \sum_i \sigma_i^c \quad \text{where} \quad i-j \geq 3 \land p_{ij} \geq 0.3
\]

\[
\sigma_i^c = p_{ij} e^{-(d_i^c - d_{ij}^c)^2} + p_{ij} \frac{d - d_i^c}{d - d_{ij}^c} \quad \text{when} \quad d_i^c < d \\
- p_{ij} e^{-(d_i^c - d_{ij}^c)^2} + p_{ij} \frac{d - d_i^c}{d - d_{ij}^c} \quad \text{when} \quad d_i^c > D \\
- p_{ij} \quad \text{otherwise}
\]

\[
\sigma^c = \sum_i \sigma_i^c \quad \text{where} \quad i-j \geq 3 \land p_{ij} \geq 0.3
\]

Algorithm 1 Constraint Guided Neighbours for PSP

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

| Type | $\phi$ range | $\psi$ range |
|------|--------------|--------------|
| G    | [-59, -39]   | [-36, -16]   |
| H    | [-67, -47]   | [-57, -37]   |
| I    | [-67, -47]   | [-80, -60]   |
| B    | [-130, -110] | [110, 130]   |

| Type | $\phi$ range | $\psi$ range |
|------|--------------|--------------|
| G    | [-130, -110] | [110, 130]   |
| H    | [-180, 180]  | [-180, 180]  |
| I    | [-180, 180]  | [-180, 180]  |
| B    | [-180, 180]  | [-180, 180]  |
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 $3^\circ$ 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 $\text{tabu}(i', j', r)$ 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 $s_{c_{ij}}$ 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 $s_{c_{ij}}$ is strictly better than $s_{c_{ij}}$. We also alternatively accept $c'$ when $s^c$ is strictly better than $s^{c'}$ and $\chi^c$ is not worse than $\chi^{c'}$. Next, steric clash minimisation is basically a secondary objective. So when steric clash based score $\chi^c$ is chosen in Line 7, we accept neighbour $c'$ with best $\chi^{c'}$ in Line 20, if $\chi^{c'}$ is strictly better than $\chi^c$ 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 $\alpha$ type, 11 $\beta$ type, and 10 $\alpha/\beta$ type proteins. These proteins are from existing PSP search algorithms such as QUARK [5], MODE-K [9], and MOD-CSA/CA [43] or a machine learning algorithm such as SPOT-ID [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 $\chi^c$, Algorithm 1 uses (i) distance map based scoring function $s^c$, (ii) tabu with tenure 15, (iii) initialisation using predictions from SPOT-ID [39], (iv) selection of residue pairs $i$ and $j$ based on scoring functions $\chi_{c_{ij}}$ or $s_{c_{ij}}$, and (v) generation of $\Delta \phi$ and $\Delta \psi$ values from the ranges determined by the MAE values of SPOT-ID. 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 $\sigma^c$ 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 $i - j \geq 3$ and $r_{ij} < 0.5$ 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 and then selects a loop residue in between residues $i$ and $j$.
- **ri**: unlike dm, initialises the $\phi$ and $\psi$ angles randomly but using the SS specific angle ranges shown in **TABLE 1**.
- **fr**: like rl, initialises the $\phi$ and $\psi$ angles randomly from the SS specific angle ranges and unlike dm, generates $\Delta \phi$ and $\Delta \psi$ values from the full range of $[-180^\circ, +180^\circ]$.

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 $\leq$ 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](image)

**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 ≤ 0.5Å. Note that 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.
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.

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    | α β  | 138    | CGNP              | 2 hrs 10 mins|
|         |      |        | CGLFOLD          | 3 hrs 20 mins|
|         |      |        | trRosettaX       | 25 mins      |

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

VIII. CONCLUSIONS

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.

ACKNOWLEDGEMENTS

This research is partially supported by Australian Research Council Discovery Grant DP180102727.

REFERENCES

[1] John Juniper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. Nature, page 1, 2021.
[2] Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhengling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences, 117(3):1496–1503, 2020.
[3] Jun Liu, Xiao-Gen Zhou, Yang Zhang, and Gui-Jun Zhang. Cglfold: a contact-assisted de novo protein structure prediction using global exploration and loop perturbation sampling algorithm. Bioinformatics, 36(8):2443–2450, 2020.
[4] Jinbo Xu. Distance-based protein folding powered by deep learning. Proceedings of the National Academy of Sciences, 116(34):16856–16865, 2019.
[5] Dong Xu and Yang Zhang. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Structure, Function, and Bioinformatics, 80(7):1715–1735, 2012.
[6] Xiao-Gen Zhou, Chun-Xiang Peng, Jun Liu, Yang Zhang, and Gui-Jun Zhang. Underestimation-assisted global-local cooperative differential evolution and the application to protein structure prediction. IEEE Transactions on Evolutionary Computation, 24(3):536–550, 2019.
[7] Shuangbao Song, Shangle Gao, Xiaochao Qian, and Yuki Todo. Aminoes: Archive information assisted multi-objective evolutionary strategy for ab initio protein structure prediction. Knowledge-Based Systems, 146:58–72, 2018.
[8] Saulo HP de Oliveira, Eleanor C Law, Jiye Shi, and Charlotte M Deane. Sequential search leads to faster, more efficient fragment-based de novo protein structure prediction. Bioinformatics, 34(7):1132–1140, 2018.
[9] Xingqian Chen, Shuangbao Song, Junkai Ji, Zheng Tang, and Yuki Todo. Incorporating a multiobjective knowledge-based energy function into differential evolution for protein structure prediction. Information Sciences, 540:69–88, 2020.
[10] Leonardo de Lima Corrêa, Bruno Borguesas, Mathias J Krause, and Márcio Dorn. Three-dimensional protein structure prediction based on memetic algorithms. Computers & Operations Research, 91:160–177, 2018.
[11] Leonardo de Lima Corrêa and Márcio Dorn. A multi-population memetic algorithm for the 3-d protein structure prediction problem. Swarm and Evolutionary Computation, 55:100677, 2020.
[12] Hong, Wenkai Su, Zong Wang, Zhengling yang Du, Shang-Hua Peng, Ming-Ming Gao, Jianyi Cheng, and Yang. Improved protein structure prediction using a new multi-scale network and homologous templates. Advanced Science, 8(24), 2021.
[13] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lemart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. CHARMM: the biomolecular simulation program. Journal of computational chemistry, 30(10):1545–1614, 2009.
[14] Tyka M Leaver-Fay A and Lewis SM. Rosetta3: an object- oriented software suite for the simulation and design of macromolecules. Methods Enzymol., 487:545–574, 2011.
[15] Zhang Jian and Yang Zhang. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS One, 5(10), 2010.
[16] J.Richmond Timothy. Solvent accessible surface area and excluded volume in proteins: analytical equations for overlapping spheres and implications for the hydrophobic effect. Molecular Biology, 138(1), 1984.
| Protein Type | Protein ID | Protein Length | RMSD | GDT |
|--------------|------------|----------------|------|-----|
| **α**        | 6SBR       | 130            | 4.347 | 0.706 | 0.718 | 0.369 |
|              | 7JQK       | 127            | 5.23  | 5.807 | 0.56  | 0.554 |
|              | 6VAX       | 113            | 7.865 | 9.148 | 0.498 | 0.419 |
| helices and loops | 6T3I | 86             | 3.843 | 5.103 | 0.61  | 0.58  |
|              | 6SIF       | 51             | 3.073 | 5.946 | 0.686 | 0.571 |
|              | 6V8H       | 141            | 7.3   | 8.632 | 0.59  | 0.547 |
|              | T0959125    | 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 and loops | T096825    | 114            | 8.316 | 13.028 | 0.388 | 0.319 |
|              | 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  |
| sheets and loops | 6UXC | 172            | 10.488 | 10.503 | 0.352 | 0.516 |
|              | T0949      | 183            | 15.865 | 19.655 | 0.183 | 0.178 |
|              | T0958      | 84             | 5.669 | 6.393 | 0.5   | 0.39  |
|              | T0997      | 185            | 8.91  | 8.926 | 0.45  | 0.446 |
|              | T1022S1    | 223            | 14.342 | 16.668 | 0.288 | 0.277 |

[30] Fereshteh Mateie moghadam, MA Hakim Newton, Abdullah Dehzangi, Abdul Karim, B Jayaram, Shoba Ranganathan, and Abdul Sattar. Enhancing protein backbone angle prediction by using simple models of deep neural networks. Scientific Reports, 10(1):1–12, 2020.

[31] Badri Adhikari. A fully open-source framework for deep learning protein real-valued distances. Scientific Reports, 10(13374), 2020.

[32] Badri Adhikari. RealDist: Real-valued protein distance prediction. 2020.

[33] Wang Sheng, Siqi, Zhen Sun, Renyu Lin, Jinbo Zhang, and Xu. Accurate de novo prediction of protein contact map by ultra-deep learning modeling. Computational Biology, 13(1), 2017.

[34] Jie Hou, Tianjui Wu, Renzhi Cao, and Jinbo Cheng. Protein tertiary structure model driven by deep learning and contact distance prediction in caspi13. Proteins: Structure, Function, and Bioinformatics, 87(12):1165–1179, 2018.

[35] Luciano A Abriata, Giorgio E Tamo, Bohdan Monastyrskyy, Andriy Kryshthafovych, and Matteo Dal Peraro. Assessment of hard target modeling in caspi12 reveals an emerging role of alignment-based contact prediction methods. Proteins: Structure, Function, and Bioinformatics, 83(8):1436–1449, 2015.

[36] Badri Adhikari and Jinliang Cheng. Confold: improved contact-driven ab initio protein folding. Proteins: Structure, Function, and Bioinformatics, 3(4):4039–4045, 2018.

[37] David T. Jones Tomasz Kosciolek. De novo structure prediction of globular proteins by deep neural networks. Bioinformatics, 34(23):4039–4045, 2018.

[38] Christophe N Magnan and Pierre Baldi. SSpro/ACCpro 5: almost perfect secondary structure predictions. Bioinformatics, 35(14):2403–2410, 2018.

[39] Jack Hanson, Kuldip Paliwal, Thomas Litfin, Yuedong Yang, and Yaoqi Zhou. 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.

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

[41] Fred Glover. Tabu search methods in artificial intelligence and operations research. ORSA Artificial Intelligence, 1(2):6, 1987.

[42] David T. Jones Tomasz Kosciolek. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. Plos One, 9(3), 2014.

[43] Jianzhu Ma, Sheng Wang, Zhiyong Wang, and Jinbo Xu. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. Bioinformatics, 31(21):3506–3513, 2015.

[44] Sheng Wang, Wei Li, Renyu Zhang, Shiwang Liu, and Jinbo Xu. Coinfold: a web server for protein contact prediction and contact-assisted protein folding. Nucleic acids research, 44(W1):W361–W366, 2016.

[45] Sheng Wang, Siqi Sun, Zhen Li, Renyu Zhang, and Jinbo Xu. Accurate de novo prediction of protein contact map by ultra-deep learning model. PLoS computational biology, 13(1):e1005324, 2017.

[46] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. Nature, 579(7792):707–710, 2020.

[47] Vincenzo Cutello, Giuseppe Narzisi, and Giuseppe Nicosia. A multi-objective evolutionary approach to the protein structure prediction problem. Journal of The Royal Society Interface, 3(6):139–151, 2005.

[48] Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. Hierarchical structure of proteins. In Molecular Cell Biology. 4th edition. WH Freeman, 2000.

[49] Yang GYuedong, Hefferran Rhys, Paliwal Kul dip, Lyons James, Dehzangi Abdullah, Alok Sharma, Wang Jihua, Sattar Abdul, and Zhou Yaoqi. Spider2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. Methods Molecular Biology, 1484:55–63, 2017.
and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 25(17):3389–3402, 1997.

RIANON ZAMAN is a PhD student at School of ICT at Griffith University, Australia. She obtained her MSc. from United International University, Bangladesh. She completed her Bsc. from Khulna University of Technology (KUET). Her research interest is Bioinformatics, Optimization and Machine Learning.

M.A.H. NEWTON is a researcher at Institute for Integrated and Intelligent Systems (IIIS), Griffith University, 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 National ICT Australia (NICTA). His research interests are in Intelligent Search, Machine Learning, and Bioinformatics.

F. MATAEIMOGHADAM is a PhD student at School of ICT at Griffith University, Australia. She has completed her masters from Kharazmi University, Iran. She pursued her Bachelor degree from Ferdowsi University of Mashhad, Iran. Her research interest is Data science pipelines for biological data, Optimization, Statistical Learning and mathematical modeling for regression, classification, and clustering.

A. SATTAAR is a professor at 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.