PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2025, **27**, 12220

Preserving structural integrity: fold reproducibility in computational design of proteins non-homologous to wild-type sequences†

Bondeepa Saikia and Anupaul Baruah *

Received 10th April 2025,
Accepted 7th May 2025

DOI: 10.1039/d5cp01373a

rsc.li/pccp

Even with remarkable accomplishments, designing a protein with a given structure is still a challenging task. There is no general approach that works for all challenges. Protein sequences with higher sequence similarity are usually shown to have similar three dimensional structures. This work is focused on designing non-homologous protein sequences with low sequence similarity to the wild-type sequence while maintaining secondary structure integrity. Basically, the aim of the present study is to check whether or not dissimilar sequences tend to encode a similar structure. In this work, we employ a negative design approach to design protein sequences by optimizing non-native conformational ensembles. Three non-native conformational ensembles are created for each of the three chosen target structures. During the design of protein sequences using the Monte Carlo simulation method and developed Cα distance-based statistical potentials, these ensembles are destabilized along with stabilization of the targets. The structures of the designed sequences are determined using AlphaFold2. Interestingly, the results suggest that secondary structure elements like alpha helices and beta sheets can be conserved even for non-homologous sequences with low sequence similarity. It is also observed that the designed sequences have the ability to reproduce the three target protein's fold *viz.* all-α, all-β and mixed αβ despite very low sequence similarity to the wild-type sequences. This indicates that the employed design strategy is effective in preserving structural integrity despite low sequence similarity.

1 Introduction

Computational protein design (CPD) aims to create energetically optimal novel sequences that reliably fold into predefined target structures. CPD is useful in determining the sequence–structure relationships.[1–3] Due to its potential to create novel biomolecules with specialized functions like catalysts,[4,5] therapeutic proteins,[6–9] biosensors[10–12] *etc.*, this field has gained significant attention. Most structure-based protein design methods aim at achieving a higher level of accuracy in sequence as well as structure recovery rates. However, this remains challenging despite many computational advancements in the field of protein design. However, different graph-based deep-learning methods such as ProteinMPNN,[13] AlphaDesign,[14] PiFold (protein inverse folding)[15] *etc.* have shown to improve the accuracy in sequence recovery.

Proteins can be designed employing many approaches *viz.* positive design, negative design, and machine learning based sequence design. Positive design approaches involve stabilization of the native state only. In the positive design approach the competing non-native states are not explicitly disfavored during the design of protein sequences for the desired target fold. Negative design, on the other hand, chooses sequences that destabilize the competing non-native conformations, guaranteeing that only the target fold is energetically favorable. Machine learning based sequence design methods such as AlphaFold2 based protein design approaches[16–18] and diffusion models[19] predict sequences for given target structures. However, unlike negative design, these approaches do not specifically account for competing non-native states and instead they primarily rely on evolutionary data. Moreover, a lower success rate in protein design can be attributed to the reliance on the positive design approaches in designing protein sequences. Thus, in order to design protein sequences that are non-homologous to wild-type sequences and are less prone to misfolding, negative design approaches are more effective in designing such sequences.

Most experimentally characterized proteins with similar sequences encode a similar type of structure.[20] However, the literature provides proof of many structurally similar proteins with low sequence similarities. Thus, ensuring the designed

Department of Chemistry, Dibrugarh University, Dibrugarh-786004, India.
E-mail: anupaulbaruah@dibru.ac.in

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5cp01373a

sequence exhibits low sequence similarity to the wild-type protein sequence while precisely reproducing the desired three dimensional structure is a quite challenging task. This challenge is specifically very crucial for *de novo* protein design, because *de novo* protein design[21,22] aims to design protein sequences with low resemblance to any existing protein databases. Therefore, this can be accomplished by carefully balancing structural integrity with sequence variation, so that the designed sequences maintain the desired three dimensional structure even while their amino acid composition diverges dramatically.

The ability to reliably reproduce a desired protein fold in novel sequences despite significant sequence variations can be termed as fold reproducibility.[23,24] Computational protein design methodologies combine in-depth understanding of protein energetics with specific scoring systems and statistical potentials in designing protein sequences for a given target structure. Over the past two decades, tremendous efforts have been made to develop efficient potentials that can capture the actual protein energetics.[25,26] In order to maintain protein fold stability, most protein design methods rely on evaluating inter-residue interactions, side-chain packing and backbone geometry.[27,28] Utilization of knowledge-based statistical potentials derived from large datasets of experimentally determined protein structures is a commonly employed approach in the protein design procedure.[29–31] Fold reproducibility in design workflows has been improved by recent developments in computational tools, such as machine learning-based techniques[32] and models based on AlphaFold.[33] These tools have been found to be incorporated into design loops while designing protein sequences.[16,17] Despite the advances in machine learning, deep learning-based methods for protein design, it still remains challenging to develop one computational method that can be used to design any protein.

The design of protein sequences that have low sequence similarity to the wild-type sequences while maintaining fold integrity has wide-ranging applications in synthetic biology, drug design, and biotechnology. Additionally, *de novo* design of proteins broadens the range of strategies available for developing novel biomolecules with specific characteristics not seen in nature. Advances in protein design approaches are opening the door to more robust and adaptable protein designs by solving the combined issues of sequence divergence and structural preservation. Thus, the ability to design protein sequences non-homologous to wild-type sequences will pave the way for a better understanding of sequence–structure relationships and will push the boundaries that can be achieved in the field of protein design.

In our previous work,[34,35] the foldability criterion Δ,[36] which incorporates negative design features, was found to be an optimal criterion for designing more stable and target compatible protein sequences. Moreover, several other studies[37,38] assert that negative design is important for protein design. Inspired by the previous work's results, this study investigates the fold reproducibility ability of the sequences designed using negative design approaches for three structurally classified proteins (an all-α, an all-β and a mixed αβ). Monte Carlo simulation in sequence space is applied to design sequences utilizing the foldability criterion Δ for three selected target structures. The threading method is applied to generate an ensemble of non-native conformations for each selected target protein. A one body and different two body statistical potentials are developed as a function of Cα distances of amino acid residues. These developed potentials are used to assess the stability of a designed sequence in a given conformation. The three dimensional structures of the designed sequences are predicted using AlphaFold2.[33] The secondary structures of proteins are typically assigned using the DSSP (define secondary structure of proteins) algorithm.[39] Here, we used DSSP3 (3-states-DSSP) to assign the secondary structures of the designed as well as the target proteins. The major goal of this work is to examine whether or not the designed sequences with low sequence similarity to the wild-type sequences encode a similar fold to the target structure. Therefore, designing protein sequences having significant sequence variation will help to assess how sequence mutations affect the overall fold of the protein. Since the method gives non-homologous sequences, it provides a scope to study a sequence space, which may be non-existent in nature and may also fold to existing structures. This opens up a whole new possibility of novel sequences. The results suggest that the designed sequences encode a similar fold to the target protein despite significant sequence variation. The secondary structure comparison between the designed and the target proteins shows that the secondary structure content found in the target proteins is well maintained in all the designed proteins. It is also found that all designed proteins preserve the residue–residue contacts found in the targets to a great extent. Despite significant sequence variations, the ability to design protein sequences that encode folds similar to the target proteins demonstrates the effectiveness of the developed statistical potentials in capturing protein energetics. Again, preservation of secondary structure content found in the target proteins highlights the effectiveness of the potentials in modeling the stabilizing forces that govern local structural motifs. Moreover, residue–residue contact conservation suggests that the developed statistical potentials capture important interactions that are essential for preserving the target proteins' overall fold. The developed statistical potentials capture protein energetics because they are derived from observed frequencies of structural features *i.e.* residue's Cα distances in 500 known protein structures. Thus, the findings of this work suggest that the developed statistical potentials are able to capture the protein energetics to a certain extent. Therefore, this work suggests that the protein design method such as this one will be quite helpful in understanding the protein sequence–structure relationship that will ultimately lead to the *de novo* design of protein sequences.

2 Methods

Protein design refers to searching of an amino acid sequence compatible with a desired target structure. The foremost step of any protein design procedure is to chose the target structure, followed by choosing of foldability criteria and most

importantly an energy function or potential to quantify the sequence energy in a given conformation.

2.1 Target structures

Three experimentally resolved protein structures with PDB ID 6FM8 (chain length = 50, X-ray resolution = 1.78 Å, *R*-value = 0.270, an all-α protein) (Fig. 1a), 2IGD (chain length = 61, X-ray resolution = 1.10 Å, *R*-value = 0.125, a mixed αβ protein) (Fig. 1b) and 1HOE (chain length = 74, X-ray resolution = 2.00 Å, *R*-value = 0.199, an all-β protein) (Fig. 1c) are chosen from the RCSB PDB (Research Collaboratory for Structural Bioinformatics Protein Data Bank). The Cα backbone of these proteins is taken as the target structure for the protein design procedure. Selection of the proteins is based on their secondary structure content, availability of high resolution structure and moderate sequence length.

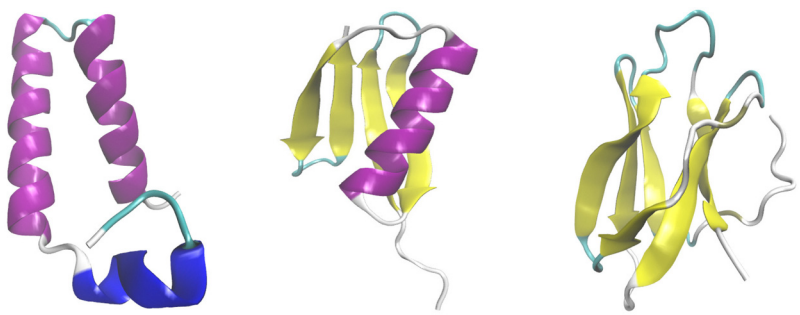
2.1.1 Non-native conformations. The utilization of negative design approaches has demonstrated their significance in enhancing the success rate of *in silico* design of protein sequences.[34,35,40,41] Stable protein sequences for a specific target can be designed by destabilizing a competing non-native conformational ensemble. Consequently, the threading method is employed to construct an ensemble of real-like non-native conformations for each of the three target structures. For this purpose, a data set of 3989 proteins are compiled from the RCSB PDB. With one chain and no polymer entities, this protein data set has a resolution better than 3 Å, a sequence identity of 30%, and all structures identified by X-ray diffraction. The data set is filtered to exclude proteins with the chain length shorter than 74, missing Cα coordinates, and missing residues. There are 556 proteins (see Table S2 of ESI†) in the final data set of proteins used to generate the ensemble of real-like non-native conformations by threading. The following is the process for generating an ensemble of non-native conformations, say for protein 1HOE (chain length = 74). To begin, we extract the Cα coordinates of all 556 proteins having chain lengths greater than equal to 74 from their corresponding PDB files. A non-native conformation with a chain length of 74 is generated by using the first 74 Cα coordinates of a protein, beginning from residue site i through $i + 73$. The subsequent conformation of the same protein thus begins at residue sites $i + 5$ through $i + 5 + 73$ since the next five sites are not included while creating the subsequent conformation. The 5-residue gap ensures that the conformations are not overly similar due to shared local structural context. By omitting the immediate 5-residue overlap, we aim to reduce structural redundancy and correlation between successive conformations. This increases the diversity of sampled non-native conformations, making them more appropriate for evaluating folding specificity and stability. Throughout the entire protein, this process is carried out to generate every conceivable real-like conformation of 74 chain length. This is done for each of the 556 proteins, and for the target 1HOE, an ensemble of 16 881 real-like non-native conformations is generated. Similarly, 18 319 non-native conformations of chain length 61 are generated for the target 2IGD and 19 536 non-native conformations of chain length 50 are generated for the target 6FM8. These three sets of conformational ensembles of the proteins 6FM8, 2IGD and 1HOE are considered for protein design procedures. Root mean square deviations (RMSDs) of these three sets of non-native conformational ensembles are determined with respect to respective targets and the mean RMSDs of these three ensembles of non-native conformations are tabulated in Table 1.

2.2 Foldability criteria

A well defined foldability criterion is a prerequisite for any protein design procedure. A foldability criterion quantifies the fitness of a sequence to a given conformation in terms of optimized potential or energy function.[42–44] In protein design, preventing a protein from adopting non-native or undesirable conformations is the major goal of negative design approaches. With this approach, the native conformation is maintained as the most stable conformation while protein sequences are designed to be energetically unfavorable for alternative conformations. Negative design in terms of stability gap, Δ,[36] has proven to be a more effective criterion for designing protein sequences.[34,35] Therefore, in this work, we have selected the foldability criterion Δ to design protein sequences. This foldability criterion can be expressed mathematically as

$$\Delta = E_f - \langle E \rangle_u \quad (1)$$

Here, the first term E_f represents energy at the target state and the second term $\langle E \rangle_u$ represents average energy over all non-native conformations.



(a) (b) (c)

Fig. 1 Three-dimensional structures of (a) 6FM8, (b) 2IGD and (c) 1HOE.

Table 1 Non-native conformational ensembles with their mean RMSD

Target	Total non-native conformations	Mean RMSD (Å)
6FM8	19 536	12.52
2IGD	18 319	13.66
1HOE	16 881	14.36

2.3 Potentials for design

Potentials or energy functions are used in protein design methods to distinguish between sequences that fit the target structure and those that do not. The energy of a sequence in a given conformation can be quantified by using potentials.[29] Two statistical potentials are developed in this work and employed in the design of protein sequences.

2.3.1 One body statistical potential. This work develops a one-body statistical potential using a dataset of 500 globular proteins compiled from the RCSB PDB subjected to certain criteria: all 500 proteins are of one chain, contain no polymer entities, have resolution better than 3 Å, have a sequence identity of 30%, and all structures are identified by X-ray diffraction. The identity, surrounding environment and position of a residue in a given protein conformation are the sole factors that affect a one-body potential. The developed potential is a function of residue's Cα distances from neighboring residues. First, the Cα distances between an amino acid residue, say α and its 10 nearest neighboring amino acid residues are calculated. These 10 nearest neighbor distances can be presented as NN_i, i = 1, 2, 3, …, 10. This is done for all such amino acid residues α present in the protein data set and the total number of occurrences, $N_t(\alpha)$, of an amino acid α in the data set is also calculated. These distances are then grouped into 200 distinct bins, $B_j j$ = 1, 2, 3, …, 200, each 0.2 Å in size. Then, the total number of Cα distances for amino acid α and for each NN_i within each bin, $N_{B_j}^{NN_i}(\alpha)$ is calculated. Then, the probabilities of a bin of amino acid α for each NN_i are calculated as

$$P_{B_j}^{NN_i}(\alpha) = \frac{N_{B_j}^{NN_i}(\alpha)}{N_t(\alpha)} \tag{2}$$

Then, for each NN_i and for each bin, B_j, the probabilities $P_{B_j}^{NN_i}(\alpha)$ for all 20 amino acids are summed as given in eqn (3) and this is followed by calculation of their averages as given in eqn (4)

$$S_{B_j}^{NN_i} = \sum_{\alpha=1}^{20} P_{B_j}^{NN_i}(\alpha) \tag{3}$$

$$A_{B_j}^{NN_i} = \frac{S_{B_j}^{NN_i}}{20} \tag{4}$$

Then the propensity of an amino acid α to be in a particular bin B_j in each NN_i can be calculated as

$$\Pr_{B_j}^{NN_i}(\alpha) = \frac{P_{B_j}^{NN_i}(\alpha)}{A_{B_j}^{NN_i}} \tag{5}$$

Finally, this propensity $\Pr_{B_j}^{NN_i}(\alpha)$ is converted to potential by applying the Boltzmann inversion method[45] as follows

$$V_{B_j}^{NN_i}(\alpha) = -k_B T \ln \Pr_{B_j}^{NN_i}(\alpha) \times P_{B_j}^{NN_i}(\alpha) \tag{6}$$

Here, k_B is the Boltzmann constant and the value of $k_B T$ is assumed to be unity.

2.3.2 Two body statistical potential. A two body potential takes into account the interaction between pairs of residues in a protein. Among the many features on which two body contact depends is the distance between Cα atoms. For the purpose of protein sequence design, two body statistical potentials are developed based on the Cα distance between amino acid residue pairs. For this, the same data set of 500 globular proteins compiled from the RCSB PDB, as mentioned in Section 2.3.1, is considered. The potentials are developed by considering different cut-offs for Cα distances and these are (0–5) Å, (5–6) Å, (6–7) Å, (7–8) Å, (8–9) Å, (9–10) Å, (10–11) Å, (11–12) Å, (12–13) Å, (13–14) Å and (14–15) Å. There are a total of 210 unique interacting amino acid pairs for 20 naturally occurring amino acids. The total number of two body contacts, N_{tot}^k, k = 1, 2, 3, …, 11, in the data set are first calculated using the stated cut off distances. Next, the number $N(\alpha_i,\alpha_j)$, which represents the number of inter-residue contacts made by each of the 210 distinct amino acid residue pairs, is determined. The probability of a unique residue pair (α_i,α_j) forming a contact is computed as

$$P_{\text{contact}}^k(\alpha_i, \alpha_j) = \frac{N(\alpha_i, \alpha_j)}{N_{tot}^k} \tag{7}$$

Then, we have determined the unbiased probability of contact formation by unique residue pairs. This is done by first calculating the probability that each amino acid in the data set is in contact. In order to accomplish this, we determine the number of times each amino acid α_i is involved in contact formation within the data set by calculating $N_{\text{contact}}(\alpha_i)$. Afterwards, the following formula is used to determine their respective probabilities of contact formation:

$$P_{\text{contact}}^k(\alpha_i) = \frac{N_{\text{contact}}(\alpha_i)}{2 \times N_{\text{tot}}^k} \tag{8}$$

Therefore, the unbiased probability of residues α_i,α_j forming a contact is computed as

$$P_{\text{unbias}}^k(\alpha_i,\alpha_j) = P_{\text{contact}}^k(\alpha_i) \times P_{\text{contact}}^k(\alpha_j) \tag{9}$$

Finally, the 11 two body statistical potentials are obtained by applying the Boltzmann inversion method[45] as follows

$$E_{2B}^k = \left[-k_B T \ln \frac{P_{\text{contact}}^k(\alpha_i, \alpha_j)}{P_{\text{unbias}}^k(\alpha_i, \alpha_j)}\right] \times P_{\text{contact}}^k(\alpha_i, \alpha_j) \tag{10}$$

2.4 Design of protein sequences

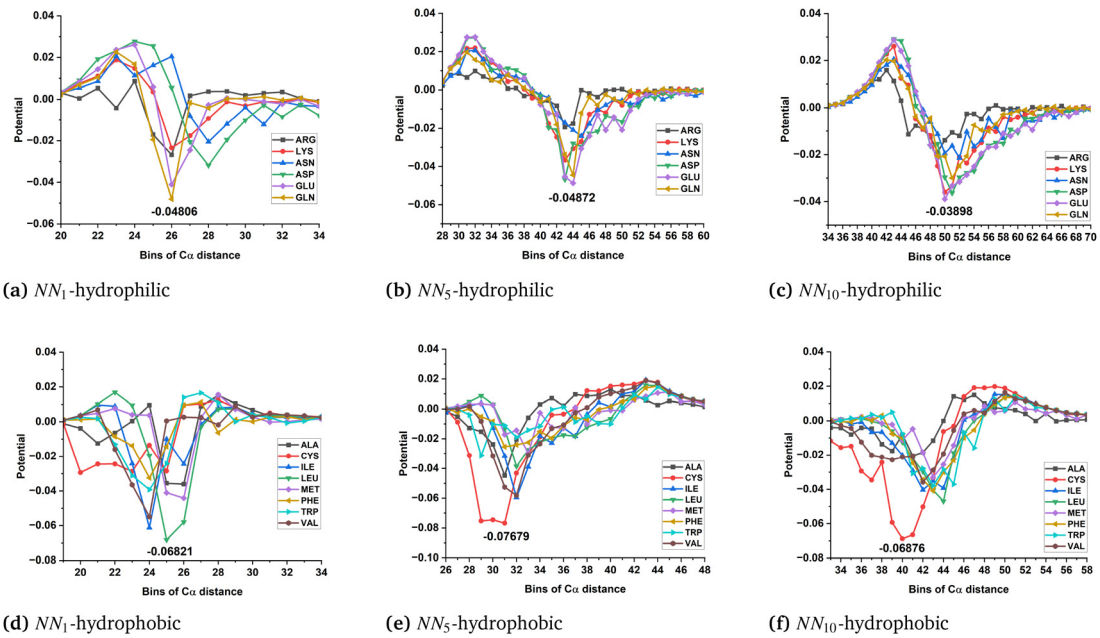
Protein sequences are designed for the target structures 6FM8, 2IGD and 1HOE using Monte Carlo simulation in sequence

space. Δ is used as the foldability criterion with a combination of the developed one body and two body statistical potentials for the purpose of protein sequence design. Three Monte Carlo simulations, each of 3×10^4 steps, are performed for the three selected target structures. Each simulation is initiated by generating a random sequence, which is considered as an input for the simulation. For this input sequence the value of foldability criterion Δ_{old} is calculated. Then, a random residue site is selected and at that site a point mutation is incorporated randomly. Then, the value of Δ_{new} of this mutated sequence is calculated. The metropolis criterion is utilized to accept or reject the mutated sequence. Using this we have designed 120 sequences for each of the target structures 6FM8, 2IGD and 1HOE.

3 Results and discussion

3.1 Potential: distinction between hydrophobic and hydrophilic amino acids

As discussed in Section 2.3.1, a one body statistical potential is developed as a function of residue's Cα distances from neighboring residues for the design of protein sequences. Based on the size, charge, and hydrophobicity an amino acid exhibits distinct structural and energetic preferences. Similarly, different amino acids show specific preferences for either the surface or core environment. These specific preferences of amino acids can be ascertained from their corresponding potential values at specific Cα distances. The lower the value of the potential, the higher the tendency of the amino acid to occupy that specific environment. In soluble, globular proteins hydrophobic amino acids tend to occupy the core of a protein while the hydrophilic amino acids tend to occupy the surface of a protein. When it comes to membrane-integral proteins, this trend differs as hydrophobic residues are frequently surface-exposed to interact with the lipid bilayer. In the context of potential values, hydrophobic amino acid residues exhibit favorable potentials (lower potential value) for shorter Cα distances. This is because amino acid residues found in the core of a protein tend to pack tightly in that environment, while hydrophilic amino acid residues exhibit favorable potentials for longer Cα distances, as they are typically solvent-exposed. In Fig. 2a–f, one body potentials are plotted for some hydrophilic amino acid residues (ARG, LYS, ASN, ASP, GLU and GLN) and some hydrophobic amino acid residues (ALA, CYS, ILE, LEU, MET, PHE, TRP and VAL) for three different nearest number distances represented by NN_i, i = 1, 5, and 10. In Fig. 2a–c, it can be observed that the hydrophilic amino acid residues exhibit favorable potentials for longer Cα distances. In Fig. 2d–f the hydrophobic amino acid residues exhibit favorable potentials for shorter Cα distances. A similar trend is observed for all remaining nearest neighbor distances and their corresponding plots of potential *vs.* bins of Cα distances are incorporated in the ESI† as Fig. S1(a)–(n). It is observed that the most favorable potential values (the steep regions of the curve) for the CYS residue gradually decrease from NN_1 to NN_4 and then again increase to NN_{10}, with NN_4 having the lowest potential value of −0.1136. Since the potential is derived from statistical observations of residue–residue distances in a dataset of proteins compiled from the RCSB PDB, the NN_4 might represent the most favorable distance for interactions involving CYS residues. The rise in potential values beyond the NN_4 could indicate decreasing interaction frequency or less favorable energetic contributions at larger distances. From these observations, it can be inferred



(a) NN_1-hydrophilic (b) NN_5-hydrophilic (c) NN_{10}-hydrophilic

(d) NN_1-hydrophobic (e) NN_5-hydrophobic (f) NN_{10}-hydrophobic

Fig. 2 Plot of potential *versus* bins of Cα distances (each bin is of 0.2 Å size) for hydrophilic and hydrophobic amino acid residues for different nearest number distances represented by NN_i, i = 1, 5, and 10.

that the developed potential allows the distinction between hydrophobic and hydrophilic amino acid residues.

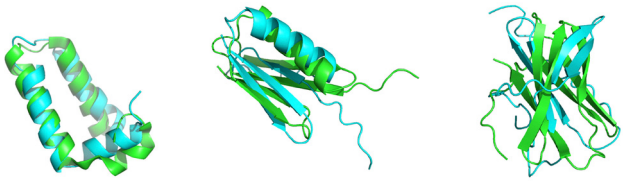
3.2 Determination of sequence similarity

Protein sequences are designed by applying Monte Carlo simulation in sequence space and the developed potentials. As discussed in Section 2.4, 120 sequences are designed for each of the three selected target proteins 6FM8, 2IGD and 1HOE respectively. These designed sequences are taken for determination of sequence similarity with their respective wild-type sequences. This is done by aligning the entire length of two sequences, *i.e.* the wild-type and the designed sequence, and the total number of exact matches between two aligned sequences is determined. The corresponding percentages of sequence similarity of all the designed sequences are also determined.

3.3 Fold reproducibility: three-dimensional structure determination

All 360 sequences (120 sequences for each target) having different sequence similarities designed for the proteins 6FM8, 2IGD and 1HOE are considered for determination of their three-dimensional structures. AlphaFold2[33] (**https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/batch/AlphaFold2_batch.ipynb**), a artificial intelligence method developed by DeepMind, is used to predict the three-dimensional structure of the designed sequences. Root mean squared deviation (RMSD) is used to compare the target protein structure and the predicted structure of the designed sequence. The RMSDs are calculated using the command ''super PDB1, PDB2'' in PyMOL.[46] The total number of aligned atoms is also determined. A sequence is selected for each target protein 6FM8, 2IGD and 1HOE based on certain criteria as given in Table 2. After subjecting these criteria on the 120 sequences designed for each of the three proteins, the predicted structures of the designed sequences having the highest number of aligned atoms are selected for further analyses. The selected designed proteins with their foldability values are given in Table 2.

In Fig. 3a–c, the top-ranked structures of these selected designed sequences predicted by AlphaFold2 are compared with their respective target structures. Here, the top-ranked structure refers to the highest-confidence model having the highest pLDDT values among the five predicted structures generated during the protein structure prediction by AlphaFold2. The full term for pLDDT is ''predicted local distance difference test''. This is a per-residue measure of local confidence used by AlphaFold and it assesses the confidence in the local structure by measuring how well the prediction agrees with an experimental structure. The RMSDs and pLDDTs (predicted local distance difference test) are also presented. In Fig. 3a, it is observed that despite very low sequence similarity to the wild-type sequence (8.00%), the sequence having a foldability value of $\Delta = -8.0582$ designed for the target 6FM8 encodes a similar fold to the target 6FM8, which can be confirmed by its RMSD value of 3.59 Å. In Fig. 3b and c, although the RMSDs are observed to be slightly greater, however, we have seen that the predicted structures of the sequences designed for 2IGD and 1HOE assume a similar type of fold *i.e.* αβ mixed and all-β as the target structures respectively. The sequence similarities of these designed sequences are also very low (Table 2). These results indicate that our developed statistical potential is able to capture the target protein characteristics in the designed sequences despite significant sequence variation. A sequence's ability to reproduce the target protein folds infers that structural stability of a protein is more dependent on the fold than on the sequence. This inference is of utmost value in protein design, where structural or functional characteristics are desired without the need for significant sequence conservation.



(a) 6FM8, RMSD=3.59 Å, pLDDT=59.5 **(b) 2IGD**, RMSD=5.26Å, pLDDT=62.7 **(c) 1HOE**, RMSD=5.91Å, pLDDT=45.0

Fig. 3 The three-dimensional structures of the target protein (colored green) and the top-ranked predicted structure of the designed protein (colored cyan) and their corresponding RMSDs and pLDDTs.

Table 2 Selection of designed sequences based on RMSD and the number of aligned atoms

	Criteria			
Target	RMSD	Aligned atoms	Foldability of selected sequence	Sequence similarity with wild-type sequence (%)
6FM8	$0 < \text{RMSD} < 4$	>200	$\Delta = -8.0582$	8.00
2IGD	$0 < \text{RMSD} < 6$	>200	$\Delta = -10.8426$	8.19
1HOE	$0 < \text{RMSD} < 6$	>200	$\Delta = -16.5262$	8.10

3.4 Secondary structure comparison

To get a better picture of the fold reproducibility ability of the designed sequences, the predicted structures of the designed sequences are compared with their respective target protein structure in the context of secondary structure contents. This is done to check whether or not the predicted structures of the designed sequences maintain the structural integrity in terms of secondary structure similarity. The same set of sequences having foldability values $\Delta = -8.0582$, $\Delta = -10.8426$ and $\Delta = -16.5262$ designed for proteins 6FM8, 2IGD and 1HOE are considered for this analysis. The secondary structures of the top-ranked predicted structures of the designed sequences as well as the target proteins are assigned using DSSP3 (3-states-DSSP). DSSP3 is a reduced representation of DSSP8 where eight types of secondary structures are assigned to amino acids. While in DSSP3, three types of secondary structures are assigned to amino acids: helix (H), strand (E), and loop (C). We have used the general reduction of 8 types of secondary structures to 3 types: (H/G/I → H, E/B → E, S/T/C → C). Fig. 4a–c are the secondary structure comparison maps between the predicted

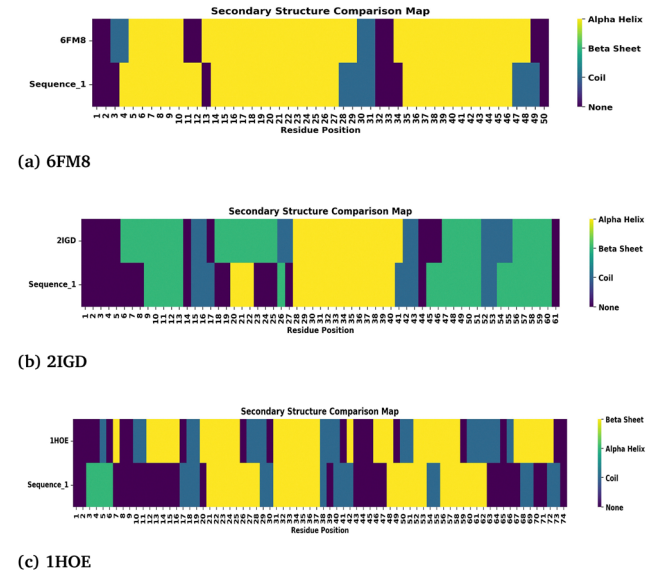


Fig. 4 Secondary structure comparison maps of the designed proteins with their respective target proteins, (a) 6FM8, (b) 2IGD, and (c) 1HOE.

structures of the designed sequences and their target structures 6FM8, 2IGD and 1HOE respectively. The percentages of secondary structure similarities are calculated for each case and are tabulated in Table 3. We have observed that the secondary structures found in the target proteins are well maintained in all the three designed sequences. It is also observed that the percentage of secondary structure similarity obtained for the sequence designed for all-α protein 6FM8 is the highest (78%). This indicates that the design approach used in this work is well-suited for all-α proteins. Nevertheless, it is also observed that the designed protein precisely captures the helical and most of the β portions of the αβ mixed protein 2IGD, and the percentage of secondary structure similarity is found to be 68.85%. This result highlights the versatility of the design approach. The versatility of the design approach is further strengthened by the percentage of secondary structure similarity of 43.24% for the sequence designed for the all-β protein 1HOE. Overall, it is seen that our employed design approach is effective in designing protein sequences that retain structural integrity despite having very low sequence similarity to the wild-type sequences. The relatively low value of secondary structure similarity for the sequence designed for all-β protein 1HOE suggests that beta-sheet-dominated structures are less tolerant to sequence changes than all-α or α dominant proteins.

3.4.1 Contact map comparison. The same set of designed sequences are then taken for analyzing whether or not they preserve the residue–residue contacts found in the target proteins. Contact maps can be used to visually represent the residue–residue contacts in protein structures. A comparison of contact maps of the designed and the target proteins is crucial in assessing how well the designed protein preserves the residue–residue contacts found in the target protein. Thus, the Cα backbones of the designed and target proteins are considered to calculate the residue–residue contacts within a cut off

Table 3 Percentage of secondary structure similarity between the top-ranked predicted structure of the designed proteins and their respective target proteins

Target	Sequence length	Total number of similarity	Percentage of similarity (%)
6FM8	50	39	78.00
2IGD	61	42	68.85
1HOE	74	32	43.24

distance of 7.5 Å and a range of 7.5 Å to 15 Å. The contact maps of the target proteins 6FM8, 2IGD and 1HOE; designed proteins and their comparison within a distance of 7.5 Å are given in Fig. 5a–c and within the range of 7.5 Å to 15 Å are presented in Fig. 6a–c. We have purposefully selected these two ranges of distances for calculating the residue–residue contacts to check how well the designed proteins capture the short-range and long-range contacts found in the targets. For comparing the contact maps, the sensitivity and specificity are calculated for all using eqn (11)

$$\left.\begin{aligned}\text{Sensitivity} &= \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \\ \text{Specificity} &= \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}}\end{aligned}\right\} \quad (11)$$

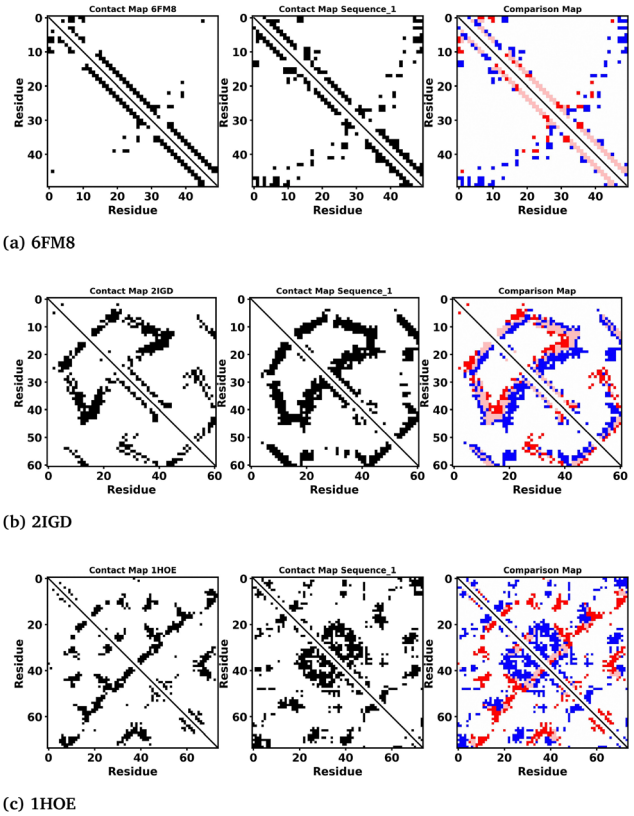
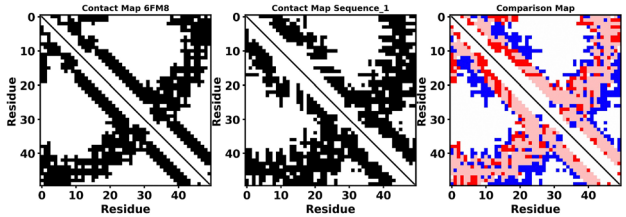
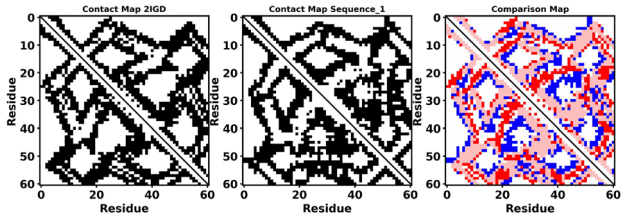


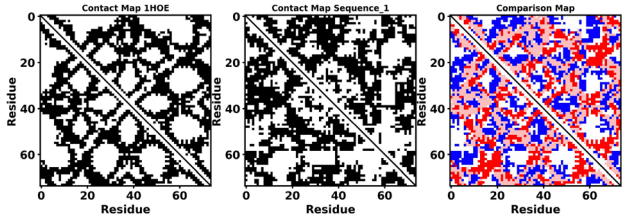
Fig. 5 Contact maps of the target protein, designed protein (sequence_1) and their comparison within a distance of 7.5 Å: (a) 6FM8, (b) 2IGD, and (c) 1HOE. In the comparison map, the contacts found in the target are colored red, the contacts found only in the designed protein are colored blue and the rose colored contacts represent the contacts present in both structures.



(a) 6FM8



(b) 2IGD



(c) 1HOE

Fig. 6 Contact maps of the target protein, designed protein (sequence_1) and their comparison within a range of distance of 7.5 Å to 15 Å, (a) 6FM8, (b) 2IGD, and (c) 1HOE. In the comparison map, the contacts found in the target are colored red, the contacts found only in the designed protein are colored blue and the rose colored contacts represent the contacts present in both structures.

Table 4 Sensitivity and specificity percentages for the residue–residue contacts found within the distance of 7.5 Å and range of 7.5 Å to 15 Å

		Sensitivity (%)		Specificity (%)	
Protein	Designed sequence	Contact < 7.5 Å	7.5 Å $\leq$ contact < 15 Å	Contact < 7.5 Å	7.5 Å $\leq$ contact < 15 Å
6FM8	$\Delta = -8.0582$	72.82	71.33	96.85	88.72
2IGD	$\Delta = -10.8426$	55.84	64.81	93.23	88.75
1HOE	$\Delta = -16.5262$	26.29	52.74	92.21	82.81

and are tabulated in Table 4. In the context of contact maps, sensitivity denotes the ability of the design method to detect the true positives and the contacts present in both designed and target proteins, while specificity measures the ability to identify the true negatives, the contacts that are absent in both proteins. We have observed a significantly high value of sensitivities and specificities in both the distance ranges for the sequence designed for the all-α protein 6FM8 and αβ mixed protein 2IGD. This ensures that the designed proteins capture both short-range and long-range residue–residue contacts found in these target proteins. The specificities are observed to be very high for both short-range and long-range residue–residue contacts for all the designed sequences. This result ensures that the design approach successfully captures the contacts absent in both designed and target proteins. Again, the sensitivity is found to be relatively low for the short-range contacts for the sequence designed for protein 1HOE; however, it is found to be significantly high for the long-range contacts. Since long-range interactions play a vital role in determining the overall fold and its stability, these results indicate that the designed proteins mimic the target structure's stability and overall fold.

A local contact map comparison is also done to check to what extent the designed proteins preserve local contacts found in the target proteins. Local contacts are related to the stability and formation of secondary structures like α-helices and β-sheets. Thus, a local contact map comparison helps to determine if designed proteins maintain the secondary structure elements of the target proteins, which is important for fold stability. For this purpose, contacts between residues that are 3 to 7 residues apart along the sequence are calculated within a cut off distance of 7.5 Å and their corresponding sensitivities (Table 5) are also determined using eqn (11). This is done for all the target and the designed proteins. Then the contact maps for the target proteins 6FM8, 2IGD and 1HOE and designed proteins and their comparison for all the designed and their target structures are plotted and presented in Fig. 7a–c. Since sensitivity measures the true positive rate, the sensitivity percentages are calculated for comparing the local contact maps of the target proteins with the designed proteins and are tabulated in Table 5. Sufficiently high sensitivity percentages are observed for the all-α protein 6FM8 (81.70%) and αβ mixed protein 2IGD (70.45%). This indicates that the designed sequences also preserve the local contacts found in their respective target proteins. However, it is found to be relatively low for the sequence designed for all-β protein 1HOE (33.87%).

High sensitivity percentages are observed in the contact map comparisons for the all-α protein 6FM8 and αβ mixed protein 2IGD in contrast to the relatively low sensitivity for all-β protein 1HOE, whether it is short-range or long-range or local residue–residue contacts. This suggests that different folds exhibit varying degrees of designability and structural flexibility. All-α and α dominant αβ mixed proteins are able to reproduce most of the residue–residue contacts because of their ability to tolerate sequence variations while maintaining structural integrity.

3.5 Ramachandran plot analysis

A Ramachandran plot is a two dimensional graphical representation of backbone dihedral angles, ϕ and ψ, of amino acid

Table 5 Sensitivity percentages for the local contacts found within the distance of 7.5 Å

Protein	Designed sequence	Sensitivity (%)
6FM8	$\Delta = -8.0582$	81.70
2IGD	$\Delta = -10.8426$	70.45
1HOE	$\Delta = -16.5262$	33.87

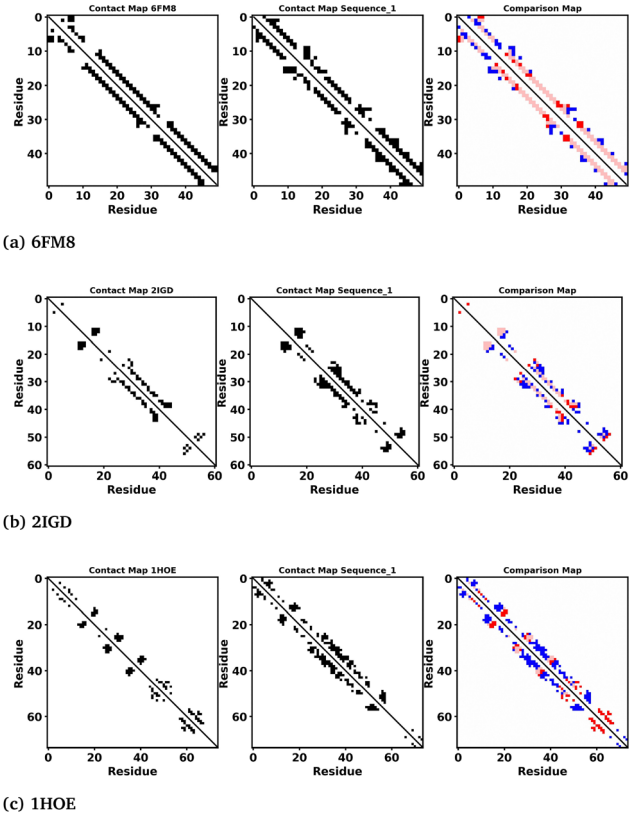
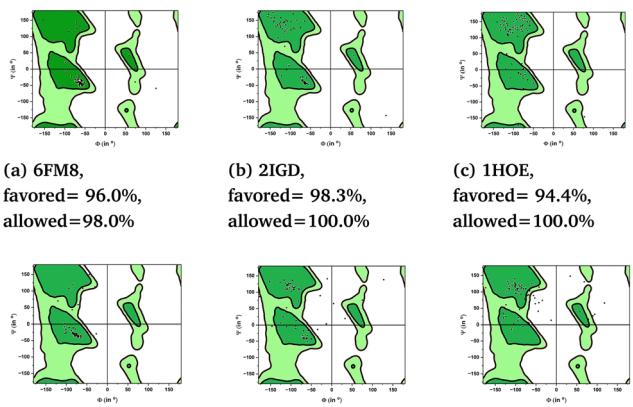


Fig. 7 Local contact maps of the target protein, designed protein (sequence_1) and their comparison within a distance of 7.5 Å, (a) 6FM8, (b) 2IGD, and (c) 1HOE. In the comparison map, the contacts found in the target are colored red, the contacts found only in the designed protein are colored blue and the rose colored contacts represent the contacts present in both structures.

residues in proteins. It can be used to analyze whether ϕ and ψ angles in the designed protein are realistic and consistent with the target proteins. Similar ϕ and ψ angle distribution in the designed protein to that of the target protein suggests that the backbone structure is well preserved in the designed protein. Thus, Ramachandran plots are constructed for all designed as well as target proteins using the data obtained from the MolProbity web server[47] and are presented in Fig. 8a–f. Fig. 8d and f exhibit a similar pattern of Ramachandran plot occupancy as in Fig. 8a–c. This means all the designed proteins exhibit similar ϕ and ψ dihedral angle distribution to that of the target proteins. This ensures that the target protein's α-helical and β-sheet portions are well-captured by all the designed proteins. All the designed proteins have most of the amino acid residues in the Ramachandran favored and allowed regions. This result assures that the backbone dihedral angles are well preserved and consistent with those of stable target protein structures.



(a) 6FM8, favored= 96.0%, allowed=98.0%
(b) 2IGD, favored= 98.3%, allowed=100.0%
(c) 1HOE, favored= 94.4%, allowed=100.0%
(d) 6FM8, Sequence_1, favored= 85.4%, allowed=97.9%
(e) 2IGD, Sequence_1, favored= 81.4%, allowed=84.7%
(f) 1HOE, Sequence_1, favored= 61.1%, allowed=79.2%

Fig. 8 Ramachandran plots for the target and the designed proteins (sequence_1) along with the percentage of residues in the favored and allowed regions.

4 Conclusions

Protein sequences with very low sequence similarity to the wild-type sequences are designed for three target structures: an all-α protein 6FM8, a mixed αβ protein 2IGD, and an all-β protein 1HOE using Monte Carlo simulation in sequence space and negative design approaches. A one body and different two body statistical potentials are developed as a function of Cα distances of amino acid residues and are used to assess the stability of the designed sequence in a given conformation. AlphaFold2 is used to predict the three-dimensional structures of the designed sequences. The results suggest that despite being non-homologous (low sequence similarity) to wild-type protein sequences, the designed protein sequences encode a similar fold to the target. The fold reproducibility ability of the designed sequences can be evaluated by analyzing how well the designed sequences preserve the secondary structure elements like α-helix, β-sheet, coil *etc.* Using DSSP3, the secondary structures of the target and AlphaFold2 top-ranked predicted structures of the designed sequences are determined. The results ensure that the structural integrity in terms of secondary structure similarity is well preserved in all the designed proteins. Contact maps are plotted for short-range and long-range contacts and are compared based on sensitivity and specificity. The results obtained from comparison of contact maps ensure that the sequence designed for the all-α protein captures both short-range and long-range residue–residue contacts found in the target protein. In the case of the sequence designed all-β protein, the sensitivity for short-range contacts is found to be relatively low as compared to sensitivities of all-α and αβ mixed proteins. However, the sensitivity for long-range contacts is found to be high for all designed proteins. This indicates that the design approach might be more promising in designing protein sequences for all-α and for α dominant αβ mixed proteins as compared to β dominant protein folds. However, the design approach efficiently captures the residue–residue short-range and long-range contacts that are absent in both the designed and target proteins very well. Comparison of local contact maps produces similar results. A comparison of our

method with RosettaDesign[48] is also carried out by designing three sequences for each target structure and is incorporated in the ESI† as Table S1 and Fig. S2–S4.

Although the structures of the designed proteins show good performance, the RMSDs and other results indicate some deviations. This deviation seems to be notable in the case of the all-β protein 1HOE. It is evident that all-β proteins frequently possess complex topologies with long loops and are structurally less regular.[26,49,50] This makes it challenging to design all-β proteins accurately than all-α and mixed αβ proteins, leading to deviations in RMSD and pLDDT values. Again, the sensitivity percentage for the residue–residue contacts within the distance of 7.5 Å for the designed all-β protein is found to be relatively less than all-α and mixed αβ proteins. In all-β proteins, short-range interactions mostly rely on exact hydrogen bonding patterns and are challenging to reproduce without atomic-level precision in design.[51] Since the design method utilized here is a coarse-grained method, there are some deviations in sensitivity for short range contacts despite the good overall performance of the designed proteins. These observed deviations may arise from methodological limitations in developed statistical potentials derived from the data set of proteins that may not fully capture the intricate physics of protein folding. The dependence on only the Cα model restricts the depiction of atomic interactions and side-chain specificity, even though the statistical potential is good at capturing coarse-grained energetic preferences. Thus, a design approach that incorporates all-atom modeling with improved potential that is capable of accurately modeling alpha helices, beta sheets, loop regions will enhance the designability and reliability of designed protein structures.

The comparison of Ramachandran plots of the designed and the target proteins exhibits a similar pattern of Ramachandran plot occupancy. This reflects that the target protein's α-helical and β-sheet portions are well captured by all the designed proteins. It is also observed that, in all the designed proteins, most of the backbone dihedral angles are well preserved and consistent with those of stable target protein structures. Thus, the findings of this work suggest that the secondary structural elements like alpha helices and beta sheets can be conserved even for non-homologous sequences with very low sequence similarity. Since the method gives non-homologous sequences, it provides a scope to study a sequence space, which may be non-existent in nature and may also fold to existing structures. This opens up a whole new possibility of novel sequences. The ability to design protein sequences with very low sequence similarity to wild-type sequences will pave the way of a better understanding of sequence–structure relationships. Therefore, a protein design method such as this one will be quite helpful in *de novo* design of protein sequences in the future.

Author contributions

Bondeepa Saikia: writing – original draft, conceptualization, methodology, software, formal analysis and Anupaul Baruah: writing – reviewing and editing, supervision, funding acquisition, conceptualization.

Data availability

The data that support the findings of this study are available within this article and its ESI.† The 3989 proteins are compiled from the RCSB PDB. The 556 proteins that are taken for generation of non-native conformations are provided in the article's ESI† as Table S2. The in-house codes used in this study are available at **https://github.com/ABLab2018/Fortran-codes**.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors gratefully acknowledge DST, India PURSE (Project No. SR/PURSE/2022/143). The authors acknowledge the financial support from DST, India (Project No. CRG/2023/008126). The financial assistance of the DST-FIST and UGC-SAP program to the Department of Chemistry, Dibrugarh University is also gratefully acknowledged.

References

1 J. G. Saven, *Curr. Opin. Colloid Interface Sci.*, 2010, **15**, 13–17.
2 I. Samish, C. M. MacDermaid, J. M. Perez-Aguilar and J. G. Saven, *Annu. Rev. Phys. Chem.*, 2011, **62**, 129–149.
3 E. Michael, S. Polydorides, T. Simonson and G. Archontis, *J. Chem. Phys.*, 2020, **153**, 054113.
4 D. N. Bolon and S. L. Mayo, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 14274–14279.
5 J. Kaplan and W. DeGrado, *Proc. Natl. Acad. Sci. U. S. A.*, 2004, **101**, 11566–11570.
6 B. Leader, Q. J. Baca and D. E. Golan, *Nat. Rev. Drug Discovery*, 2008, **7**, 21–39.
7 A. Chevalier, D.-A. Silva, G. J. Rocklin, D. R. Hicks, R. Vergara, P. Murapa, S. M. Bernard, L. Zhang, K.-H. Lam and G. Yao, *et al.*, *Nature*, 2017, **550**, 74–79.
8 L. Cao, I. Goreshnik, B. Coventry, J. B. Case, L. Miller, L. Kozodoy, R. E. Chen, L. Carter, A. C. Walls and Y.-J. Park, *et al.*, *Science*, 2020, **370**, 426–431.
9 S. B. Ebrahimi and D. Samanta, *Nat. Commun.*, 2023, **14**, 2411.
10 L. L. Looger, M. A. Dwyer, J. J. Smith and H. W. Hellinga, *Nature*, 2003, **423**, 185–190.
11 C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. W. Nelson, A. Schena, W. Jankowski, C. G. Kalodimos, K. Johnsson and B. L. Stoddard, *et al.*, *Nature*, 2013, **501**, 212–216.
12 M. J. Bick, P. J. Greisen, K. J. Morey, M. S. Antunes, D. La, B. Sankaran, L. Reymond, K. Johnsson, J. I. Medford and D. Baker, *eLife*, 2017, **6**, e28909.

13 J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas and N. Bethel, *et al.*, *Science*, 2022, **378**, 49–56.
14 Z. Gao, C. Tan and S. Z. Li, *arXiv*, 2022, preprint, arXiv:2202.01079, DOI: **10.48550/arXiv.2202.01079**.
15 Z. Gao, C. Tan, P. Chacón and S. Z. Li, *arXiv*, 2022, preprint, arXiv:2209.12643, DOI: **10.48550/arXiv.2209.12643**.
16 C. A. Goverde, B. Wolf, H. Khakzad, S. Rosset and B. E. Correia, *Protein Sci.*, 2023, **32**, e4653.
17 L. Moffat, J. G. Greener and D. T. Jones, *bioRxiv*, 2021, preprint, DOI: **10.1101/2021.08.24.457549**.
18 M. Jendrusch, J. O. Korbel and S. K. Sadiq, *bioRxiv*, 2021, preprint, DOI: **10.1101/2021.10.11.463937**.
19 J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte and L. F. Milles, *et al.*, *Nature*, 2023, **620**, 1089–1100.
20 A. Šali, J. Overington, M. Johnson and T. Blundell, *Trends Biochem. Sci.*, 1990, **15**, 235–240.
21 P.-S. Huang, S. E. Boyken and D. Baker, *Nature*, 2016, **537**, 320–327.
22 I. V. Korendovych and W. F. DeGrado, *Q. Rev. Biophys.*, 2020, **53**, e3.
23 B. Rost, R. Schneider and C. Sander, *J. Mol. Biol.*, 1997, **270**, 471–480.
24 D. J. Rigden, *From protein structure to function with bioinformatics*, Springer, 2009, vol. 355.
25 A. Goldenzweig and S. J. Fleishman, *Annu. Rev. Biochem.*, 2018, **87**, 105–129.
26 X. Pan and T. Kortemme, *J. Biol. Chem.*, 2021, **296**, 100558.
27 M. M. Gromiha and S. Selvaraj, *Prog. Biophys. Mol. Biol.*, 2004, **86**, 235–277.
28 B. Kuhlman and P. Bradley, *Nat. Rev. Mol. Cell Biol.*, 2019, **20**, 681–697.
29 J. U. Bowie, R. Lüthy and D. Eisenberg, *Science*, 1991, **253**, 164–170.
30 M. J. Sippl, *Curr. Opin. Struct. Biol.*, 1995, **5**, 229–235.
31 S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, *Chem. Rev.*, 2016, **116**, 7898–7936.
32 L. Wei and Q. Zou, *Int. J. Mol. Sci.*, 2016, **17**, 2118.
33 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek and A. Potapenko, *et al.*, *Nature*, 2021, **596**, 583–589.
34 B. Saikia, C. R. Gogoi, A. Rahman and A. Baruah, *J. Chem. Phys.*, 2021, **155**, 144102.
35 B. Saikia and A. Baruah, *Soft Matter*, 2024, **20**, 3283–3298.
36 N. V. Dokholyan and E. I. Shakhnovich, *J. Mol. Biol.*, 2001, **312**, 289–307.
37 I. N. Berezovsky, K. B. Zeldovich and E. I. Shakhnovich, *PLoS Comput. Biol.*, 2007, **3**, e52.
38 A. Rossi, C. Micheletti, F. Seno and A. Maritan, *Biophys. J.*, 2001, **80**, 480–490.
39 W. Kabsch and C. Sander, *Biopolymers*, 1983, **22**, 2577–2637.
40 C. M. Summa, M. M. Rosenblatt, J.-K. Hong, J. D. Lear and W. F. DeGrado, *J. Mol. Biol.*, 2002, **321**, 923–938.
41 W. Jin, O. Kambara, H. Sasakawa, A. Tamura and S. Takada, *Structure*, 2003, **11**, 581–590.
42 J. G. Saven, *Chem. Rev.*, 2001, **101**, 3113–3130.
43 P. Biswas, J. Zou and J. G. Saven, *J. Chem. Phys.*, 2005, **123**, 154908.
44 A. Baruah and P. Biswas, *J. Chem. Phys.*, 2015, **142**, 05B608_1.
45 D. Reith, M. Pütz and F. Müller-Plathe, *J. Comput. Chem.*, 2003, **24**, 1624–1636.
46 L. Schrödinger, The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC, 2015.
47 I. W. Davis, A. Leaver-Fay, V. B. Chen, J. N. Block, G. J. Kapral, X. Wang, L. W. Murray, W. B. Arendall III, J. Snoeyink and J. S. Richardson, *et al.*, *Nucleic Acids Res.*, 2007, **35**, W375–W383.
48 Y. Liu and B. Kuhlman, *Nucleic Acids Res.*, 2006, **34**, W235–W238.
49 E. Marcos and D.-A. Silva, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2018, **8**, e1374.
50 Y. Zhang, Y. Liu, Z. Ma, M. Li, C. Xu and H. Gong, *bioRxiv*, 2024, preprint, DOI: **10.1101/2024.10.05.616664**.
51 D. L. Minor Jr and P. S. Kim, *Nature*, 1994, **367**, 660–663.

