Received: 30 August 2021 | Revised: 11 January 2022 | Accepted: 28 January 2022

DOI: 10.1002/prot.26321

RESEARCH ARTICLE

Can the jigsaw puzzle model of protein folding re-assemble a hydrophobic core?

Gargi Biswas[1,2] | Semanti Ghosh[1] | Sankar Basu[1] | Dhananjay Bhattacharyya[1,2] | Alok Kumar Datta[3] | Rahul Banerjee[1,2]

[1]Saha Institute of Nuclear Physics, Kolkata, India

[2]Homi Bhabha National Institute, Mumbai, India

[3]Indian Institute of Chemical Biology, Kolkata, India

Correspondence
Semanti Ghosh and Rahul Banerjee, Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Kolkata 700064, India.
Email: semanti2007@gmail.com and rahul.banerjee@saha.ac.in

Present address
Semanti Ghosh, Swami Vivekananda University, Kolkata, India

Sankar Basu, Asutosh College (Under University of Calcutta), Kolkata, India

Abstract

According to the "jigsaw puzzle" model of protein folding, the isomorphism between sequence and structure is substantially determined by the specific geometry of side-chain interactions, within the protein interior. In this work, we have attempted to predict the hydrophobic core of cyclophilin (LdCyp) from *Leishmania donovani*, utilizing a surface complementarity function, which selects for high goodness of fit between hydrophobic side-chain surfaces, rather in the manner of assembling a three-dimensional jigsaw puzzle. The computational core prediction method implemented here has been tried on two distinct scenarios, on the LdCyp polypeptide chain with native non-core residues and all core residues initially set to alanine, on a poly-glycine polypeptide chain. Molecular dynamics simulations appeared to indicate partial destabilization of the two designed sequences. However, experimental characterization of the designed sequences by circular dichroism (CD) spectroscopy and denaturant (GdmCl) induced unfolding, demonstrated disordered proteins. Stepwise reconstruction of the designed cores by cumulative sequential mutations identified the specific mutation (M122L) as primarily responsible for fold collapse and all design objectives were achieved upon rectifying this mutation. In summary, the study demonstrates regions of the core to contain highly specific (jigsaw puzzle-like) interactions sensitive to any perturbations and a predictive algorithm to identify such regions. A mutation within the core has been identified which exercises an inordinate influence on the global fold, reminiscent of metamorphic proteins. In addition, the computational procedure could predict substantial regions of the core (given main-chain coordinates) without any reference to non-core residues.

KEYWORDS

Cyclophilin, hydrophobic core, jigsaw puzzle model, protein design, protein stability

1 | INTRODUCTION

The presence of a well-packed hydrophobic core segregated from the surrounding aqueous environment is a ubiquitous feature of globular proteins.[1,2] Hydrophobic residues constituting the core exhibit a much higher degree of conservation relative to surface-exposed amino acids,[3] and despite some measure of plasticity,[4] mutations within the core of native proteins are generally destabilizing.[5,6] Specific side-chain interactions within the core confer thermal stability,[7–10] influence binding specificities[10,11] or protein function,[12] and contribute to the conformational uniqueness of native proteins,[11] where uniqueness refers to the significant energy difference between the native low energy state and other alternative conformations.

Two models concerning protein cores, the "jigsaw puzzle"[12] and the "oil drop",[13] lie on the opposite ends of the spectrum and have provided quite a few insights, probably limited by their respective points of view. The jigsaw puzzle model postulates that the isomorphism between sequence and structure is to be substantially determined by the specific geometry of inter-digitating side-chains of hydrophobic residues in the protein interior. Experiments have repeatedly demonstrated that the simple conservation of volume or composition of core residues, does not assure elegant packing in a thermally stable protein.[14,15] Calculations have confirmed that at least a subset of side-chain interactions (within a core) involve non-random stereospecific geometry, exhibiting appreciable surface complementarity between their respective surfaces, somewhat akin to a three-dimensional jigsaw puzzle.[9] The jigsaw puzzle model has also been applied with some measure of success in the identification of recurring packing motifs within proteins,[10] in the inverse protein folding problem,[14] and structure validation.[15] In contrast, the liquid drop model focuses on the distribution of hydrophobicities in natively folded proteins. First, proposed by Kauzmann,[16] the model has been further refined as the "fuzzy oil drop" model[17] and is based on the understanding that the dominant force in protein folding is the hydrophobic effect involving the sequestration of hydrophobic residues from the surrounding aqueous environment, resulting in a regular gradation of hydrophobicities from the interior to the surface, of a correctly folded protein. The fuzzy oil drop model has been extensively applied to structurally rationalize the hydrophobic cores of antifreeze proteins along with their mutants,[18] immunoglobulins,[19] and in general to elucidate the influence of water as an external force field on protein structure.[17]

Although specific side-chain packing interactions in the core definitely contribute to protein thermal stability, their role in the acquisition of the native fold has hitherto been denied.[11,20] Several experiments and theoretical considerations appeared to support the view that protein fold acquisition is primarily due to the overall distribution or pattern of hydrophobic/hydrophilic residues embedded in the primary sequence.[21] However, subsequent experiments have demonstrated that the mutation of even a single strategically positioned residue in the core can dramatically alter the overall global fold in natural and designed proteins.[22,23] This leads to the concept of "metamorphic" proteins which can exist in multiple folds, on occasion triggered by minimal mutations of residues which could possibly be tipping points responsible for fold switching.[22,24,25] All this goes to show that our understanding of protein cores, especially with regard to the highly context-dependent contribution of core residues in the stabilization of protein fold, structure, and stability, could be further refined.

Despite such lacunae in our basic understanding, much has been achieved in core prediction and currently, effective repacking of the core is considered to be one of the benchmarks of protein design software.[23,26–30] SCWRL4.0[31] and other related programs[32–34] have achieved notable success in solving for the side-chain χ angles (especially in the case of core residues), given the primary sequence and main-chain coordinates. With native coordinates of the main chain and non-core side-chain atoms as input, Lazar et al. successfully recovered (native-like) cores of several trial proteins,[23,26] while the redesigned core(s) of the streptococcal protein G β1 domain characterized by NMR spectroscopy, exhibited well-ordered structures (especially when high constraints were applied on packing specificity).[27] In a similar genre of core design computations, the DESIGNER algorithm was successful in recovering cores almost identical to the native sequence.[35] Another approach to predict the core is to treat it as a sub-problem in the de novo design of an entire protein, in which only main chain coordinates of the target fold are input. De novo initiatives have successfully generated compact viable cores for coiled coils,[36] helix bundles,[37] a zinc finger motif,[38] and even for small novel protein folds.[39] The most widely used design software Rosetta Design[40] (which utilizes a Monte Carlo search procedure coupled to simulated annealing to identify amino acids in the specified positions) was used with striking success in the core design of the 4-helix bundle (105 residues) CheA phosphotransferase, where the aggressive use of the Rosetta software led to proteins with heightened thermostability.[41]

Despite such notable achievements, all the currently available computational approaches do not appear to generate viable cores (leading to thermostable proteins) as a matter of course, nor does success in a small compact fold guarantee success in larger and more complex folds (with extended cores). For a set of 108 proteins, Rosetta correctly predicted 51% of the core residues (and 27% of all residues) with respect to the native protein sequence[42] and it is almost certain that out of this set of 108 predicted proteins, not all will achieve the design target. In a detailed biophysical characterization of nine redesigned globular (monomeric) proteins by Rosetta, one was found to be completely unfolded, three aggregated, and two thermally destabilized.[43] Thus, in the course of design, it is not uncommon to generate molten cores,[39,44] attain folds quite other than the target,[45,46] and success has been reported to be achieved either by fine re-tuning of force field parameters[39] or going through successive iterations of the design cycle.[36]

In general core prediction consists of three parts a) navigation of an extensive core (sequence) space and identification of prospective core sequences b) determination of their three-dimensional side-chain conformations (χ angles) in the protein structure, and, finally c) selection of the most optimal core sequence(s) based on an appropriate function. One of the primary objectives in this work has been to test the predictive power of the jigsaw puzzle model in recapturing the extended core of cyclophilin from *Leishmania donovani* (LdCyp) (Figure 1) with a computational selection procedure highly biased in favor of high surface complementarity between side-chain surfaces of core residues. The sequence space of a 20-residue core (of LdCyp), gives rise to an astronomical 6^{20} possibilities, in case six hydrophobic residues (alanine-A, valine-V, leucine-L, isoleucine-I, phenylalanine-F, and methionine-M) are considered (as possible substitutions) at each residue position. This practically infinite space has been searched for prospective core sequences, with a novel computational procedure based on a network representation of the core. In every instance, SCWRL4.0 has been utilized to build the three-dimensional

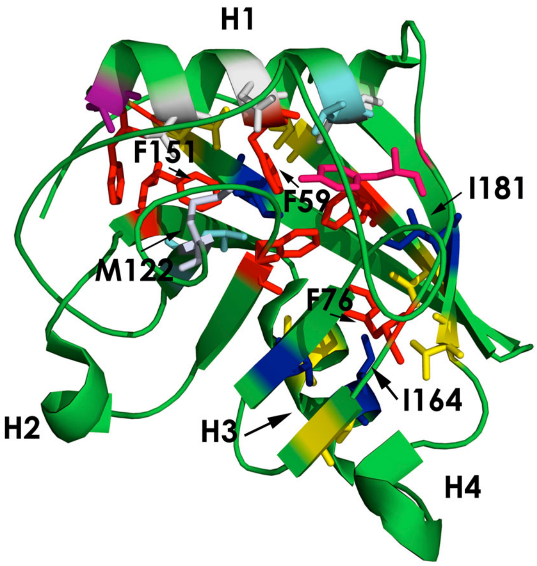


FIGURE 1 Cartoon representation of the structure of LdCyp (PDB ID: 2HAQ) which consists of a centrally located eight-stranded β-barrel surrounded by four helices (namely H1, H2, H3, and H4). The 24 hydrophobic core residues have been depicted in "stick" with different colors (F—Red, I—Blue, L—Cyan, V—Yellow, A—Violet, M—Light blue, Others—Gray). Some of the important residues (namely F59, F151, I181, F76, and M122) have been indicated by arrows

conformation of all the protein side-chains, including core or non-core residues. The optimal solution has been selected based on a surface complementarity function, which has been extensively used in the laboratory to estimate the surface fit between amino acid side-chains, and finally, the obtained core variants have been experimentally characterized for (native) fold acquisition and stability. Thus, in summary, this work has been performed to test the power of the jigsaw puzzle model in predicting core residues (of cyclophilin: LdCyp) and also to estimate the intrinsic information contained in the core, in order to self-assemble. Second, since side-chain packing defects abound within proteins, to predict if possible which regions of the hydrophobic cluster/network of core residues would be likely to resemble a three-dimensional jigsaw puzzle.

2 | MATERIAL AND METHODS

2.1 | Computational Methods

2.1.1 | Network representation of the hydrophobic core of LdCyp

All computations in this work have been performed utilizing coordinates of the high resolution (1.97 Å) crystal structure[47] of cyclophilin (LdCyp) from *L. donovani* (PDB Code: 2HAQ). The unique hydrophobic core of LdCyp, was earlier identified by visual inspection[47] of the crystal structure and calculation of the side-chain solvent accessible surface areas (SAA: probe radius 1.4 Å) of the core residues.[48] The ratio of side-chain SAA of a residue in the protein to that of an identical residue in a fully extended G-X-G polypeptide fragment (X-any amino acid residue) provides information with regard to the extent of burial (with the ratio being less than 0.05 for complete burial, Table S1). All the 24 residues (Figure 1) constituting the hydrophobic core of the molecule were found to be completely buried (burial of SAA <0.05) and can be enumerated as V29, F31, V33, I45, L47, F48, A52, T55, N58, F59, L62, C63, Y71, F76, V79, I85, L120, M122, F134, F151, V161, I164, V179, and I181.[47] Of the 24 residues T55, N58, C63, and Y71 are not "hydrophobic" amino acids (though found within the extended hydrophobic cluster) and were included in the calculations without any change in their identities. Thus, the computational core prediction procedure was concerned only with the remaining 20 hydrophobic amino acids.

The positions of the 20 hydrophobic core residues in the three-dimensional structure of cyclophilin could be considered as 20 sites on which to computationally build side-chain conformers (rotamers) of hydrophobic amino acids (A, V, L, I, F, and M; all greater than 1.80 and 0.620 in the Kyte & Doolittle and Eisenberg scales, respectively,[49,50] Table S1). Henceforth, the term "site" will refer to the three-dimensional positions of the core residues and will be specified by their residue numbers in the native polypeptide chain. In the first step of the calculation, all the conformers/rotamers from Richardson's rotamer library[51] for the hydrophobic residues (A, V, L, I, F, M) were computationally built into each of the 20 sites by the repeated application of the "fourth atom fixing procedure".[52] The (fourth atom fixing) procedure utilized the idealized bond length and bond angles from the CHARMM force field[53] and all side-chain torsion angles from Richardson's rotamer library.[51] The side-chain atoms of each such modeled conformer(s) were then tested for short contacts (less than 2.80 Å) with main-chain atoms including the side-chain C_β atom, from the rest of the polypeptide chain. A particular hydrophobic residue was "disallowed" at a specific (core) site if the side-chain atoms of all its conformers (from the rotamer library) were involved in (at least one) short contact(s) with the main-chain atoms (including C_β). In all subsequent calculations, only the "allowed" hydrophobic residues were utilized at the 20 hydrophobic core sites. All the six hydrophobic residues (A, V, L, I, F, and M) were allowed in most of the sites with the exceptions of V29, V33, L62, V79, I85 (F disallowed), and A52 (with only A and V allowed).

In the next step, a network based on probable side-chain interactions between all the 24 core residues was computed to represent the (presumptive) hydrophobic cluster. The flow chart for network generation has been described in Figure 2A. In the network (Figure 2B), each node represents a core site. Two sites were selected at a time (amounting to a total of $24 \times 23/2! = 276$ pairs) and all possible side-chain rotamer combinations of the ("allowed") residues were constructed at the two selected sites. The computational procedure to build conformers at specific sites was identical to the one described in the previous paragraph. For every such (rotamer) combination (at the two selected sites), contacts were calculated between the two sets of side-chain atoms (of the constructed rotamers), and in case of at least one contact less than 3.80 Å, the two sites were said to have an

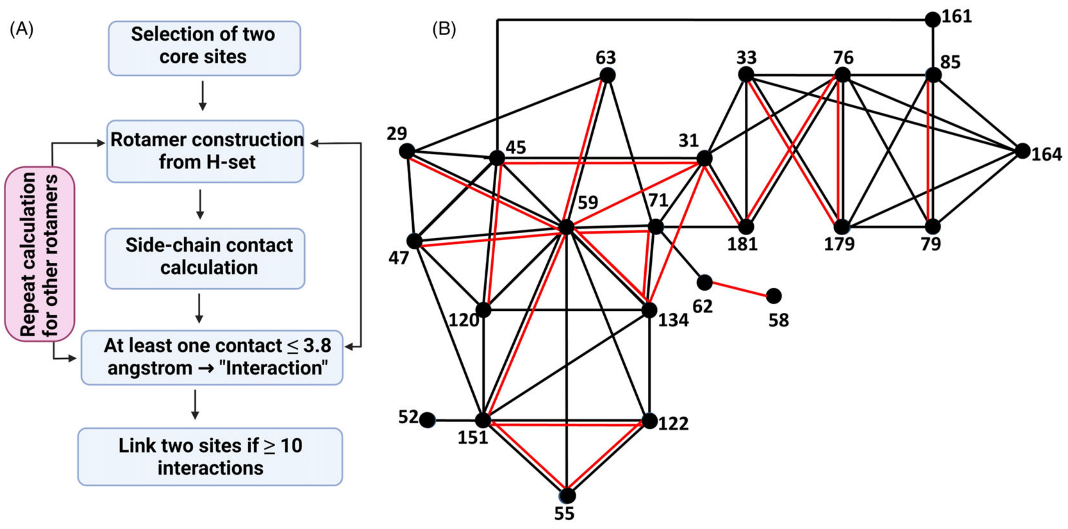

(A)
Selection of two core sites
Rotamer construction from H-set
Side-chain contact calculation
At least one contact ≤ 3.8 angstrom → "Interaction"
Link two sites if ≥ 10 interactions
Repeat calculation for other rotamers
(B)
161
63
33
76
85
29
45
31
164
59
71
47
181
179
79
62
58
120
134
52
151
122
55

FIGURE 2 Hydrophobic contact network of LdCyp. (A) The steps followed to generate the hydrophobic contact network, (B) The generated network is represented by black circles and black lines, where, the nodes (black circles) represent hydrophobic core residues or core sites. Two selected nodes were connected by a link/edge (black lines), when 10 or more rotamer combinations could be found (see Section 2) which had at least one interatomic side-chain contact less than 3.80 Å between the two sets of atoms. The red lines are from the contact map of the LdCyp crystal structure (2HAQ), two nodes are connected by an edge when any two side-chain atoms corresponding to the residues are within 3.80 Å

"interaction." If greater than (or equal to) 10 such interactions were found between the two sites, they were connected by a link (edge, indicated by black lines in Figure 2B). In other words, two nodes were connected by a link when 10 such rotamer side-chain combination pairs could be found which had at least one contact (less than 3.80 Å) between the two sets of atoms (Figure 2B). Residues at 48 and 58 positions did not satisfy the criteria to be included in the network. Residues F48 and N58 lie on the periphery of the core and their respective side-chains tend to incline away from the core centroid, even though they are completely buried by virtue of their non-core neighborhoods P23, V25, F116, and S100, I101, F106, respectively.

Reconstruction of the network based on actual side-chain contacts present in the crystal structure (2HAQ), wherein two nodes were connected by a link when the side-chain atoms of two residues had at least one contact less than equal to 3.8 Å (indicated by red lines in Figure 2B) showed that the majority of the links (17 out of 20) were actually contained in the predicted graph (described above) and only three links (31–59, 31–134, and 62–58) were exclusive to the network based on the crystal structure. These links did not satisfy the cutoff criteria (at least 10 rotamer combinations with contacts less than 3.80 Å) for inclusion in the predicted graph. All subsequent core prediction calculations were based on this (Figure 2B, Table 1) network representation of the core. Unless explicitly stated otherwise all computer programs were written locally.

2.1.2 | Protein surface generation and surface complementarity function

The generation of protein surfaces and computation of the surface complementarity function has been described in sufficient detail in numerous previous publications[54–57] from the laboratory. Prior to the generation of the Van der Waal's surface for the entire polypeptide chain all hydrogen atoms were included, whose coordinates were geometrically fixed by the program Reduce.[58] The atomic radii were assigned from the all-atom molecular mechanics force field[59] and the surface of the polypeptide chain was sampled at 10 dots/Å^2. The Van der Waals surface was generated in such a manner that the identities of individual residues constituting the polypeptide chain were preserved.[54] Thus, the surface of the entire polypeptide was sampled as an array of discrete area elements, each area element is characterized by its location (x,y,z) and the direction cosines of its normal (**dl**, **dm**, **dn**).

The surface complementarity (S_m) of side-chains corresponding to core residues was estimated by an adapted version of the function proposed by Lawrence and Colman,[60] which has previously been used in several studies.[54–57] Briefly, consider a set of dot surface points contributed by side-chain atoms of a core residue (referred to as a target). For each side-chain surface point of the target, its nearest neighbor was identified from (surface points) the rest of the polypeptide chain, within a distance of 3.5 Å. Then, following Lawrence and Colman[60]:

$$S(a,b) = \boldsymbol{n_a}.\boldsymbol{n_b} e^{-w.d_{ab}^2} \tag{1}$$

where $\boldsymbol{n_a}$, $\boldsymbol{n_b}$ are the normals to the surface point located on the target side-chain atom and its nearest neighbor (from the rest of the polypeptide chain), respectively, d_{ab}^2 is the distance between them and w a scaling constant set to 0.5. The median of the distribution of {$S(a,b)$} corresponding to all side-chain surface points of the target (which has found a neighbor to within 3.5 Å) is the surface complementarity measure S_m of the targeted core residue. The highest value attainable by

TABLE 1 The adjacency matrix for the hydrophobic core network diagram given in Figure 2. The row and column indices indicate the hydrophobic core sites and the values in the matrix are the number of rotamer combinations built at two selected sites, which have at least one interatomic contact less than 3.80 Å, between the two sets of atoms. Replacement of all numbers greater than zero by one will lead to the formal adjacency matrix in graph theory. Residues 48 and 58 did not satisfy the criteria to be included in the network

	29	31	33	45	47	52	55	59	62	63	71	76	79	85	120	122	134	151	161	164	179	181
29	–	–	–	160	68	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
31	–	–	264	69	–	–	–	–	–	12	160	28	–	–	–	–	–	–	–	–	–	302
33	–	264	–	–	–	–	–	–	–	–	–	72	–	–	–	–	–	–	–	188	153	136
45	160	69	–	–	117	–	–	54	–	–	–	–	–	–	300	–	–	–	16	–	–	–
47	68	–	–	117	–	–	–	100	–	–	–	–	–	–	30	–	–	21	–	–	–	–
52	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	15	–	–	–	–
55	–	–	–	–	–	–	–	28	–	–	–	–	–	–	–	113	–	100	–	–	–	–
59	50	–	–	54	100	–	28	–	–	16	22	–	–	–	111	75	135	115	–	–	–	–
62	–	–	–	–	–	–	–	–	–	–	95	–	–	–	–	–	–	–	–	–	–	–
63	17	12	–	–	–	–	–	16	–	–	51	–	–	–	–	–	–	–	–	–	–	–
71	–	160	–	–	–	–	–	22	95	51	–	–	–	–	–	–	24	–	–	–	–	192
76	–	28	72	–	–	–	–	–	–	–	–	–	68	306	–	–	–	–	–	148	240	88
79	–	–	–	–	–	–	–	–	–	–	–	68	–	173	–	–	–	–	–	50	36	–
85	–	–	–	–	–	–	–	–	–	–	–	306	173	–	–	–	–	–	84	144	10	–
120	–	–	–	300	30	–	–	111	–	–	–	–	–	–	–	–	63	42	–	–	–	–
122	–	–	–	–	–	–	113	75	–	–	–	–	–	–	–	–	58	65	–	–	–	–
134	–	–	–	–	–	–	–	135	–	–	24	–	–	–	63	58	–	13	–	–	–	–
151	–	–	–	–	21	15	100	115	–	–	–	–	–	–	42	65	13	–	–	–	–	–
161	–	–	–	16	–	–	–	–	–	–	–	–	–	84	–	–	–	–	–	–	–	–
164	–	–	188	–	–	–	–	–	–	–	–	148	50	144	–	–	–	–	–	–	215	–
179	–	–	153	–	–	–	–	–	–	–	–	240	36	–	–	–	–	–	–	215	–	–
181	–	302	136	–	–	–	–	–	–	–	192	94	–	–	–	–	–	–	–	–	–	–

S_m is 1.00 denoting perfect fit between two surfaces. Subsequent to the identification of all the nearest neighbors, the side-chain surface points of the target can also be partitioned into two sets whose nearest neighbors are contributed exclusively by either side-chain (N^{SC}) or main-chain atoms (N^{mC}), respectively, and the surface complementarity S_m can be estimated separately for each set. Thus, for a target, S_m^{SC} (computed from N^{SC} surface points) is an estimate of its surface complementarity with respect to neighboring side-chain atoms alone (contributed by the rest of the polypeptide chain). Since the jigsaw puzzle model emphasizes the specificity in packing between the side-chains of amino acid residues (which in any case dominates packing in protein interiors) a function SN^{SC} was defined as the product of $S_m^{SC}$and N^{SC},

$$SN^{sc} = S_m^{SC}.N^{SC} \tag{2}$$

and used as the most basic function to discriminate both the degree and extent of surface complementarity spread over the surface of a core (target) residue with respect to its immediate neighborhood constituted of side-chain atoms alone. $S_m^{SC}$was multiplied by N^{SC} as high surface complementarity spread over a wider surface area is definitely more significant than high surface fit estimates based on the chance conjunction of a few stray surface points.

2.1.3 | Hydrophobic core generation

Central to the computational process for predicting or reconfiguring the hydrophobic core of cyclophilin is the (presumptive) network representation of the hydrophobic cluster described above (Figure 2B). The core generation process considered two distinct scenarios at ascending levels of difficulty.

1. In the first case (Level 1), all the non-core native residues were included from the very start of the calculation, while all the core residues were initially set to alanine (A). In other words, the solution was obtained on a chimerical sequence consisting of native non-core residues and an altered core. With the progress of the computation (to be described below) all the 20 core sites (initially set to A) were gradually re-substituted by one of the hydrophobic residues (A, V, L, I, F, and M), based on the core prediction process described below. The initial (solved) hydrophobic cluster ("raw core"), was then optimized (see below) prior to experimental characterization.
2. In the second case (Level 2), all the residues (both core and non-core) were initially set to glycine (G), and the prospective core generated on a poly-G (non-core) polypeptide chain. This was to assess the information contained exclusively within the hydrophobic cluster to generate a viable core, without any reference to non-core residues. Subsequent to obtaining a "raw core," all non-core native residues were included and following core optimization, the obtained sequence was subject to experimental characterization. The basic computational steps (given below) to generate the (non-optimized) "raw core," was identical for both Levels (1 and 2).

To begin, the node with the highest degree (number of links) was first identified from the network diagram (Figure 2B) which is the site corresponding to residue number 59 of native LdCyp. The neighbors to this node connected by an edge in the network are 134, 120, 122, 151, 47, 45, and 29. Since the core prediction algorithm does not include within its ambit the identities of residues T55, Y71, N58, and C63, the links 59–63, 59–55, and 59–71 were not considered at this stage in the calculation. Then, on the relevant polypeptide chain (either corresponding to Level 1 or 2 obtained by appropriately modifying the coordinates 2HAQ) each pair of sites (59–134, 59–120, 59–122, etc.) were taken one at a time and SCWRL4.0[61] used to construct two side-chains (one at each site) from the hydrophobic set of residues (A, V, L, I, F, M-with due consideration of "allowed" residues at each site). Thus, there can be possible, a maximum of (6x6) 36 individual combinations (provided all hydrophobic residues are allowed at both sites), for each pair of sites (59–151, 59–120, 59–122, etc.). The program Reduce was then employed to fix hydrogen atoms on both side-chains prior to generating the protein surface of the entire polypeptide chain. This was followed by calculating the SN^{SC} for both side-chains and finally estimating their average (<SN^{SC}>). The pair of hydrophobic residues with the highest <SN^{SC}> from all side-chains pairs was selected as the solution.

The above calculation determined the highest <SN^{SC}> for F59-F151 (for both levels: Table 2), which were accordingly included in the set of predicted residues. In the next step, the network was searched for all nodes which had the highest number of links to both 59 and 151 simultaneously. Obviously, the maximum number of links can only be 2 (linked to both sites 59, 151) and these nodes were found to be (corresponding to residues) 120, 134, 122, and 47. Since this condition was satisfied for 55 as well, the solution set was expanded to (F59, F151, and T55). The above calculation was then repeated individually for the four tetrads, namely, F59-F151-T55-134, F59-F151-T55-120, F59-F151-T55-122, and F59-F151-T55-47, and the highest <SN^{SC}> determined (averaged over all the 4 residues in the tetrads), subsequent to rebuilding their side-chains afresh with SCWRL4.0 and hydrogen fixation by Reduce. For each tetrad, there could be a maximum of 6 possibilities (or less in case of disallowed residues) as there could be 6 residue substitutions for the newly included node (e.g., F59-F151-T55-A134, F59-F151-T55-V134, F59-F151-T55-L134, etc.). The highest <SN^{SC}> (for both Levels 1 and 2) was found to be F at site 122 and the solution set was consequently expanded to include (F59, F151, T55, F122). Successive cycles of the calculation designed above led to the generation of the entire core, the rules of which can be succinctly described in the following steps.

A node was considered in the next iterative step of the computation when it had the highest number of links to the cluster determined from the previous step. In case of degeneracy, that is for more than one node satisfying this condition the calculation was repeated separately for all the (prospective) nodes. For each unique residue combination, the side-chain atoms of all the residues in the cluster were rebuilt afresh by SCWRL4.0 and hydrogen atoms fixed by Reduce. SN^{SC} was then averaged over all the residues in the prospective cluster including the residue at the newly included node and the highest $<SN^{SC}>$ accepted as the solution. As has been mentioned previously, in case anyone of the nodes corresponding to residues 55, 63, 71, and 58 satisfied the condition of having the maximum number of links to a cluster, they were included in the list of solutions, retaining their native amino acid identities. The iterative application of these rules generated the "raw" core(s) (Table 3). The sequence in which core residues have been included in the computational prediction process defines a path and the paths for Levels 1 and 2 have been depicted in Figure 3. Several rounds of calculation showed a very high propensity of methionine to be included as a solution. Hence, to hinder the formation of a poly-M core, two additional criteria were applied for the inclusion of methionine in the core, namely S_m^{SC}, and the fraction surface points used for the estimation of S_m^{SC}(with respect to the total number of side-chain surface points of the targeted M: N^{SC}/N_{tot}) would individually have to exceed 0.50 and 0.75, respectively, in addition to having the highest $<SN^{SC}>$ in a cycle, to be accepted as a solution at a specific site.

TABLE 2 Highest $<SN^{SC}>$ values involving core residue/site 59 and linked sites as depicted in Figure 2. The results clearly show that for both Level 1 (ALL_R) and Level 2 (181_R) the optimal solution is for 59F-151F, based on the maximum value of $<SN^{SC}>$

181_R			ALL_R		
Residue at 59	Residue at linked site	$<SN^{SC}>$	Residue at 59	Residue at linked site	$<SN^{SC}>$
F	151 F	95.0	F	151 F	364.8
L	122 F	34.4	F	122 F	209.7
I	134 I	47.6	F	134 F	156.7
F	120 L	65.3	F	120 F	259.0
I	47 L	87.8	I	47 L	260.1
F	45 V	50.0	F	45 F	200.8
L	29 L	27.0	L	29 F	210.4

TABLE 3 The predicted raw, converged, and optimized hydrophobic core sequences (see Main Text) for Level 1 (ALL) and Level 2 (181). The sequence identities with respect to the native LdCyp and calculated core volumes for each predicted core sequence have been included in the Table. All residues which differ from the native sequence have been highlighted

Residue number	Native	ALL_R	ALL_C	ALL_O	181_R	181_C	181_O
Sequence identity (%)		50	55	65	40	55	70
Core volume (Å^3)	3257.8	3412.8	3409.0	3309.2	3391.6	3409.3	3286.3
59	F	F	F	F	F	F	F
151	F	F	F	F	F	F	F
122	M	***F***	***F***	***L***	***F***	***F***	***L***
134	F	F	F	F	F	F	F
120	L	***F***	***F***	L	L	***F***	L
47	L	L	L	L	L	L	L
45	I	***L***	***L***	***V***	I	***V***	***V***
29	V	***I***	***I***	***I***	***I***	***I***	***I***
31	F	F	F	F	F	F	F
76	F	F	F	F	***L***	***V***	***V***
181	I	***V***	***V***	***V***	***F***	***F***	***F***
33	V	V	V	V	***L***	V	V
179	V	***I***	***I***	***I***	***I***	V	V
164	I	I	***M***	***M***	***L***	I	I
79	V	***I***	V	V	***A***	***I***	V
85	I	***V***	I	I	***L***	I	I
161	V	***I***	***I***	V	***F***	***F***	V
52	A	***V***	***V***	***V***	***V***	***V***	***V***
62	L	L	L	L	L	L	L
48	F	F	F	F	***I***	F	F

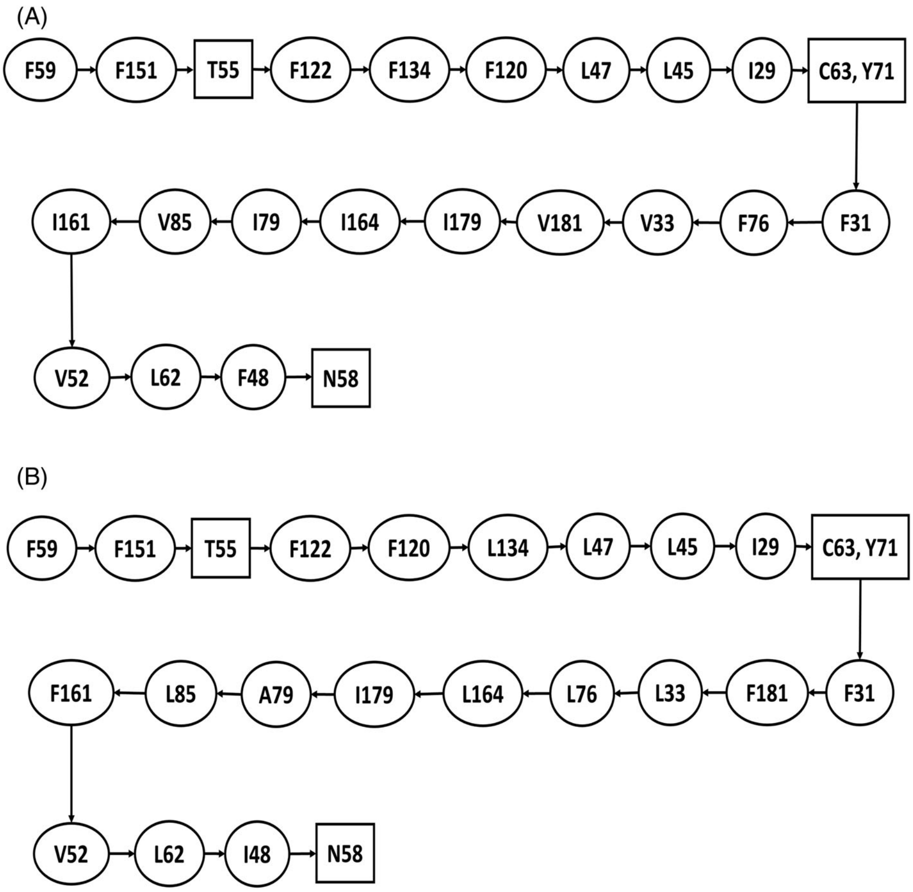


FIGURE 3 The sequence in which predicted residues were included, to re-assemble the initial "raw" (see Main Text) core during the computational prediction process for (A) Level 1 and (B) Level 2. Notably, the predicted residues were identical up to F31, for both Levels 1, 2 and diverge thereafter

The raw cores obtained for both levels were then made to converge. This consisted in selecting a site from the 20 predicted residues and recalculating $<SN^{SC}>$, for all allowed residues in the hydrophobic set, substituted at that specific site. The residue with the highest value of $<SN^{SC}>$ was then (either retained) or substituted at that site. For Level 2 the convergence process was performed twice. The raw core was initially converged on the poly-G (non-core) polypeptide chain followed by re-convergence on a chimerical sequence consisting of native non-core residues with an alternative core. Both Levels were converged by selecting residues in the same order in which the solutions (raw core) were obtained till no further change was observed in the list of predicted residues. The core volumes of the final converged solutions were estimated by the summation of the individual side-chain volumes of residues constituting the core[62] (Table 3).

2.1.4 | Core optimization

In all the predicted solutions of the protein core, a systematic over estimation of the core volume was observed (Table 3). Thus, in addition to these solutions, further optimization of the alternative cores was deemed necessary. This essentially consisted of the step-wise controlled reduction of core volume while maintaining as far as possible optimal packing (within the given volume constraints). It was observed (though with some notable exceptions) that those sites which consistently substituted larger residues in the core (in terms of volume with respect to the native protein) were followed by amino acids of lower volume, generally within a very narrow range of $<SN^{SC}>$ values. In other words, the solution at these sites exhibited a "wobble" implying the reduced margin (in terms of $<SN^{SC}>$) of the preferred residue relative to other amino acids. Therefore, at each site, the difference between the highest value of $<SN^{SC}>$ and the next lower values ($\Delta<SN^{SC}>$) was calculated and in case this difference was within 3.0 (arbitrary units) the residue with the least volume was selected to replace the residue determined previously (Tables 4 and 5). The raw, converged, and optimized core sequences obtained for Level 1 and 2 have been named ALL_R, ALL_C, and ALL_O and 181_R, 181_C, and 181_O, respectively.

2.2 | Molecular Dynamics Simulations

All atom molecular dynamics (MD) simulations for mutated cyclophilin polypeptide chains and those with reconfigured hydrophobic cores were conducted using the GROMACS (GROningen MAchine for Chemical Simulations) package,[63] version 2019.1. The native cyclophilin coordinates were obtained from the crystal structure of cyclophilin (PDB Code: 2HAQ). The mutated amino acid residues of the newly designed proteins were individually changed to their respective residues using the "build and edit protein" tool of the

TABLE 4 Optimization of ALL_C to yield ALL_O. Δ <SN^{SC}> refers to the difference in <SN^{SC}> between the highest and the second-highest value. In case Δ<SN^{SC}> was within 3.0 arbitrary units and included a residue of lower side-chain volume, it was selected for that site. The final optimized solution ALL_O is given in the column on the extreme right. NA refers to "disallowed" residues (see Section 2) at that site

	<SN^{SC}>						
Native	**First Position**	**Second Position**	**Third Position**	**Fourth Position**	**5th position**	**Δ<SN^{SC}>**	**Selected residue**
59 F	F-389.34	I-374.57	L-372.21	V-368.38	A-365.16	14.76	59 F
151 F	F-389.34	V-367.86	A-367.78	I-360.69	L-353.54	21.48	151 F
122 M	F-389.34	L-386.70	I-386.57	V-383.25	A-375.28	1.91	122 L
134 F	F-389.34	L-379.73	I-377.98	V-377.32	A-367.76	9.60	134 F
120 L	F-388.34	L-387.77	V-385.58	I-383.25	A-377.12	1.57	120 L
47 L	L-389.34	F-382.52	V-380.93	I-380.69	A-365.08	6.82	47 L
45 I	L-389.34	V-389.21	I-387.57	A-381.56	F-369.49	0.12	45 V
29 V	I-389.34	V-384.56	L-377.59	A-368.33	NA	4.77	29 I
31 F	F-389.34	L-379.60	I-377.99	V-371.15	A-362.96	9.74	31 F
181 I	V-389.34	I-387.32	L-385.66	A-383.72	F-368.97	2.02	181 V
33 V	V-389.34	I-387.15	L-377.83	A-377.52	NA	2.19	33 V
76 F	F-389.34	I-368.20	V-363.31	A-357.48	L-348.05	21.13	76 F
164 I	M-389.34	L-375.85	F-374.38	I-366.24	A-365.29	13.49	164 M
179 V	I-389.34	L-384.58	V-383.86	A-366.50	F-347.49	4.76	179 I
79 V	V-389.34	I-385.79	L-383.39	A-380.77	NA	3.55	79 V
85 I	I-389.34	L-378.09	V-377.20	A-367.08	NA	11.25	85 I
161 V	I-389.34	V- 386.96	F-386.69	L-384.00	A-374.94	2.43	161 V
52 A	V-389.34	A-385.61	NA	NA	NA	3.72	52 V
48 F	F-389.34	L-379.47	I-378.56	V-377.58	A-375.40	9.87	48 F
62 L	L-389.34	I-384.46	V-381.42	A-370.90	NA	4.88	62 L

Discovery Studio 2.5 package. The mutant proteins were minimized using the smart minimizer algorithm after applying the CHARMm force field[53] in Discovery Studio 2.5 software suite. All atoms of the system were assigned by the CHARMm 27 force field[53] and the TIP3P water model (transferable intermolecular potential with 3 points) was applied to solvate the molecule (native or variant cyclophilin coordinates) in a cubic box maintaining a minimum distance of 1 nm from the cube edges. Counter ions (Na^+ and Cl^-) were added to the solvent to keep the system charge neutral at pH 7.0. Then allowing for a 2 fs time step, all bond lengths were constrained using the Linear Constraint Solver (LINCS) method,[64] and long-range electrostatic interactions were calculated utilizing the smooth Particle Mesh Ewald (PME) method.[65] Initially, steepest descent minimization was performed with a tolerance value of 1000 kJ/mol/nm to relieve steric clashes and inappropriate geometry. After several computational trials, it was decided that thermal stability for each mutant or alternative core was best characterized by studying the thermal unfolding of the (native and) altered protein by means of MD simulations. Initially, the native protein was subject to simulation trials at 310, 335, 360, 385, 410, and 460 K, and analyses of these individual trajectories revealed that for the native enzyme, unfolding was initiated at 385 K, while complete unraveling of the protein was demonstrated at 410 K. Thus, for every cyclophilin variant, MD simulations were performed at three different temperatures: 310, 385, and 410 K, which effectively scanned the unfolding process for the native and altered proteins. Subsequent to energy minimization, the system was equilibrated (at the targeted temperature 310 or 385 or 410 K) using position-restrained (PR) MD in two steps. First, 100 ps of isochoric-isothermal (NVT) equilibration was performed using the Berendsen thermostat[66] at the selected temperature, (i.e. 310 K or 385 K or 410 K). This was followed by a simulation of 500 ps (at the same temperature, 1 bar of atmospheric pressure) in an isothermal-isobaric (NPT) ensemble utilizing the Parrinello-Rahman barostat.[67] The well-equilibrated system was then used to run the production MD for 200 ns in three replicates with trajectory frames saved every 2 picoseconds and leap-frog integration in time steps of 2 fs. The coulombic and Van der Waals short-range distance cut-off was set to 10 Å. Generated MD trajectories were analyzed primarily by calculating the root mean square deviations (RMSD) of the backbone atoms of the proteins and root mean square fluctuations (RMSF) of their C_α atoms. Analysis was performed using GROMACS in-built programs and also by locally generated software.

2.3 | Protein expression and purification

All previous experiments performed in the laboratory involving His-tagged LdCyp, have been on the protein expressed from the gene

TABLE 5 Optimization of 181_C to yield 181_O. Δ<SN^{SC}> refers to the difference in <SN^{SC}> between the highest and the second-highest value. In case Δ<SN^{SC}> was within 3.0 arbitrary units and included a residue of lower side-chain volume, it was selected for that site. The final optimized solution 181_O is given in the column on the extreme right. NA refers to "disallowed" residues (see Section 2) at that site

	<SN^{SC}>						
Native	**First Position**	**Second Position**	**Third Position**	**Fourth Position**	**Fifth position**	**Δ<SN^{SC}>**	**Selected-residue**
59 F	F-389.34	I-374.57	L-372.21	V-368.38	A-365.16	14.76	59F
151 F	F-389.34	V-367.86	A-367.78	I-360.69	L-353.54	21.48	151F
122 M	F-389.34	L-386.70	I-386.57	V-383.25	A-375.28	1.91	122 L
134 F	F-389.34	L-379.73	I-377.98	V-377.32	A-367.76	9.60	134F
120 L	F-388.34	L-387.77	V-385.58	I-383.25	A-377.12	1.57	120 L
47 L	L-389.34	F-382.52	V-380.93	I-380.69	A-365.08	6.82	47 L
45 I	L-389.34	V-389.21	I-387.57	A-381.56	F-369.49	0.12	45 V
29 V	I-389.34	V-384.56	L-377.59	A-368.33	NA	4.77	29I
31 F	F-389.34	L-379.60	I-377.99	V-371.15	A-362.96	9.74	31F
181 I	V-389.34	I-387.32	L-385.66	A-383.72	F-368.97	2.02	181 V
33 V	V-389.34	I-387.15	L-377.83	A-377.52	NA	2.19	33 V
76 F	F-389.34	I-368.20	V-363.31	A-357.48	L-348.05	21.13	76F
164 I	M-389.34	L-375.85	F-374.38	I-366.24	A-365.29	13.49	164 M
179 V	I-389.34	L-384.58	V-383.86	A-366.50	F-347.49	4.76	179I
79 V	V-389.34	I-385.79	L-383.39	A-380.77	NA	3.55	79 V
85 I	I-389.34	L-378.09	V-377.20	A-367.08	NA	11.25	85I
161 V	I-389.34	V- 386.96	F-386.69	L-384.00	A-374.94	2.43	161 V
52 A	V-389.34	A- 385.61	NA	NA	NA	3.72	52 V
48 F	F-389.34	L-379.47	I-378.56	V-377.58	A-375.40	9.87	48F
62 L	L-389.34	I-384.46	V-381.42	A-370.90	NA	4.88	62 L

inserted in a pQE32 vector.[47,68] Thus, in order to obtain altered sequences associated with the design process; stepwise, single (M122 L, I164 M, V29I, A52 V, and I181F), double (181_O2, ALL_O2, ALL_O2'), triple (ALL_O3, ALL_O3'), and quadruple mutations (ALL_O4 and ALL_O4') were performed beginning with the native (LdCyp gene) sequence (in the pQE32 vector). The details of the nomenclature associated with the sequences can be found in Table 6. These mutant sequences were obtained from KPC Life Sciences Private Limited (Kolkata, India). In addition, several designed sequences (181_O, ALL_O, ALL_O5, ALL_O6, 181_O3, 181_O4, and 181_O5) were also synthesized and inserted into the pET28a vector (obtained from Genscript, New Jersey, USA). To confirm that the expressed protein from both vectors gave identical experimental results-expression, circular dichroism, and GdmCl-mediated unfolding of the native LdCyp from both vectors (pQE32, pET28a) were performed. The almost identical experimental results confirmed the vector independent behavior of the LdCyp transcript (Figure S1, Table S2). In addition, ALL_O4' transcript has been checked in both pET28a and pQE32 vectors which resulted in almost identical expression profiles, CD traces, and denaturant induced unfolding patterns confirming the vector independent behavior of the other transcripts as well (data not shown).

The purification of mutants and core variants in the PQE32 plasmid was very similar to native LdCyp described in detail in several previous reports.[47,68–70] Mutant sequences in the pQE32 and pET28a (+) plasmids were transformed into *E. coli* M15 and *E. coli* BL21 cell lines, respectively. The altered sequences were induced with 1 mM IPTG at 20°C followed by purification on a Ni-NTA column (Qiagen, Germany). Subsequent to purification the proteins were extensively dialyzed in a buffer containing 20 mM Tris–HCl, 20 mM NaCl and 0.02% NaN_3 (pH −8.5) for storage. For spectroscopic experiments, the proteins were again dialyzed in a buffer containing 20 mM potassium phosphate (pH −7.5). Analytical grade reagents and chemicals used for protein purification and other experiments were obtained from Sisco Research Laboratories (Mumbai, India) while media used for bacterial growth was bought from Himedia (Mumbai, India).

2.4 | Circular dichroism & fluorescence spectroscopy

The far-UV CD spectra of the cyclophilin mutants and core variants were measured on a JASCO J815 (Maryland, USA) spectrophotometer in the wavelength range of 200–250 nm. 15 μM concentration of the protein (cyclophilin variants) were measured separately in a rectangular quartz cell of 1 mm path length, at 1 nm data interval and 50 nm/min scan speed. The measured molar ellipticity values (θ_λ) were converted to mean residue ellipticity (MRE) values with the formula,

TABLE 6 Nomenclature for the mutants and core variants during the sequential reconstruction of the optimized sequences (ALL_O and 181_O)

Sequence name	Mutations
ALL_O	V29I + I164 M + M122 L + A52 V + V179I + I181 V + I45 V
ALL_O2′	V29I + A52 V
ALL_O2	V29I + I164 M
ALL_O3	V29I + I164 M + A52 V
ALL_O4′	V29I + I164 M + M122 L + A52 V
ALL_O3′	V29I + I164 M + M122 L
ALL_O4	V29I + I164 M + A52 V + V179I
ALL_O5	V29I + I164 M + A52 V + V179I + I181 V
ALL_O6	V29I + I164 M + A52 V + V179I + I181 V + I45 V
181_O	I181F + F76V + A52 V + V29I + I45 V + M122 L
181_O2	I181F + F76V
181_O3	I181F + F76V + A52 V
181_O4	I181F + F76V + A52 V + V29I
181_O5	I181F + F76V + A52 V + V29I + I45 V

$$\mathrm{MRE} = \frac{M\theta_\lambda}{10dcr} \tag{3}$$

where M is the molecular weight of the protein in Da, θ_λ is molar ellipticity in millidegree, d path length in cm, c protein concentration in mg/ml, and r the total number of amino acid residues in the polypeptide chain.

The denaturant-induced unfolding of the proteins was monitored separately by intrinsic tryptophan fluorescence on a Hitachi F7000 spectrofluorometer (Illinois, USA) with a 10-μM concentration of each protein. Each of the proteins was pre-incubated overnight at varying concentrations of GdmCl (0–2.5 M with increments of 0.1 M) and the spectra were measured in a quartz cuvette of 1 cm path length, keeping both the excitation and emission band-pass filters to 5 nm. The spectra were recorded by exciting the samples at 295 nm and by measuring the emission spectra in the wavelength range 310 to 450 nm. The observed fluorescence intensities have been converted to the relative intensities for a set of data using the formula (F_{obs}-F_{min})/(F_{max}-F_{min}), where $\mathrm{F_{obs}}$ is the observed fluorescence intensity and F_{min} and F_{max} are the minimum and maximum intensities respectively, observed in a particular set of data. All experiments were repeated at least thrice.

The GdmCl mediated unfolding of LdCyp has been studied previously[70] in our laboratory and following precedent the unfolding curve was optimally fitted to a three-state unfolding equation. In an attempt to fit the unfolding curves of mutants and core variants optimally, the three-state unfolding equation gave the best fit compared to the two-state unfolding equation (**Figure S2**). The three-state unfolding equation is given by,

$$S_{\mathrm{obs}} = \frac{S_N + S_D e^{-\Delta G_{ND}/RT} + S_I e^{-\Delta G_{NI}/RT}}{1 + e^{-\Delta G_{ND}/RT} + e^{-\Delta G_{NI}/RT}} \tag{4}$$

where S_{obs} is the observed fluorescence intensity, and S_{N}, S_{D}, and S_{I} are the intensities for native, intermediate, and completely denatured protein, respectively, ΔG_{ND} and ΔG_{NI} are the difference in Gibb's free energy during N↔D and N↔I process while R, T being the ideal gas constant and absolute temperature, respectively. Moreover, ΔG_{ND} and ΔG_{NI} are linearly dependent upon the denaturant concentration and can be expressed as,

$$\Delta G_{ND} = \Delta G_{ND}(\mathrm{H_2O}) - m_{\mathrm{ND}}[D] \tag{5}$$

$$\Delta G_{NI} = \Delta G_{NI}(\mathrm{H_2O}) - m_{\mathrm{NI}}[D] \tag{6}$$

where ΔG_{ND} ($\mathrm{H_2O}$) and ΔG_{NI} ($\mathrm{H_2O}$) are the free-energies of unfolding in the absence of any denaturant, [D] denaturant concentration, while m_{ND} and m_{NI} are the slope of denaturation process N↔D and N↔I, respectively. The transition mid-points [C_{m}(N↔D) and C_{m}(N↔I)] are the [D] values for which ΔG_{ND} and ΔG_{NI} are zero.

3 | RESULTS

3.1 | Generation of alternative hydrophobic cores

LdCyp primarily consists of an eight-stranded β-barrel with two α-helices (H1, H3) at either end (Figure 1). The only extended hydrophobic core of LdCyp is situated primarily within the barrel with 19 of its 24 residues located on β-strands, 3 on α-helices H1, H3, and the remaining 2 on inter-connecting loops. The cyclophilin core appears to be close to optimal as it is highly conserved across several species spanning the animal, plant, and microbial kingdoms (Table S3). Although the core forms one continuous cluster, it can be divided into two parts—a densely packed region centered about F59 (as a hub) and a relatively loosely packed cluster around F76. The two parts of the core are conjoined at a narrow bottleneck constituted of residues F31 and I181. F59 lies on α-helix H1 and its strategic protrusion into the β-barrel makes it the primary hub of the hydrophobic cluster (Figures 1 and 2B). Most of the three-dimensional features of the cyclophilin core (derived from the crystal structure) are to a large extent captured by its network representation (Figure 2B), where each node represents a core site and is linked to a neighbor when there are a sufficient number of contacts between rotamer substitutions (from the hydrophobic set A, V, L, I, F, and M) at the two sites.

In the first round of calculations (Level 1), all non-core native residues were retained and all the 20 core sites were initially set to alanine. Governed by a minimal set of rules (see Section 2) and beginning with the node of the highest degree (corresponding to 59, Figure 2B), residues were substituted (one at a time), at each of the 20 core sites. The amino acid substitutions in the core generation process were based on the values of a surface complementarity function (<SN^{SC}>, See Section 2), which was selected for extensive and optimal fit, between side-chain surfaces of residues constituting the core. The stepwise addition of residues to generate the "raw" core (ALL_R) has been depicted in Figure 3. The process was initiated by the unambiguous prediction of F59-F151 (conserved in the native sequence) whose <SN^{SC}> value (Table 2) was significantly higher relative to all other pairs (59–134, 59–122, 59–120, etc.). However, the computational procedure generally tended to over-pack the core (Table 3), borne out by substitutions F122 (M in native LdCyp) and F120 (L in native). The total side-chain volume of the native and the raw core ALL_R were 3257.8 and 3412.8 Å^3, respectively. All the sites which had phenylalanine in the native sequence (59, 151, 134, 76, 31, 48) were conserved in the predicted core, and divergence from the native sequence was pronounced after 33. Subsequent to predicting all the 20 core residues (raw core), each core site or node was taken one at a time and the calculation was repeated by substituting (allowed) hydrophobic residues at that site, with the rest of the core sites retaining their current residue identities. For this convergence process, the residues were selected in the same order in which the raw cores were derived. <SN^{SC}> (averaged over all the 24 residues) of the raw core was 376.7 and rose to 389.3 upon the conclusion of the iterative convergence procedure. The convergence was to a core sequence (ALL_C) which differed from ALL_R only in three sites (164I → M, 79I → V, and 85 V → I). Next, an attempt was made to optimize the core by the controlled reduction in volume while maintaining comparable packing, due to the systematic overpacking observed in the process of core generation (Table 3). This was by substituting residues of lower volume at a site, in case, its <SN^{SC}> was within 3.0 (arbitrary units), of the most preferred solution. This led to changes 122F → M, 120F → L, 45 L → V, 161I → V (in ALL_C) to generate the alternative core ALL_O (Table 4). Additional constraints in surface complementarity and overlap (see Section 2) had been imposed on methionine residues, which it failed to satisfy, M122 (in ALL_O) was set to L as the next best residue. Thus, the final sequence (ALL_O) differed from the native sequence in 7 sites, with a reduced core volume of 3309.2 Å^3.

In the next round of calculations (Level 2), all the residues (both core and non-core) of the LdCyp polypeptide chain were set to glycine, and an identical computational procedure was used to generate a raw core. Core generation on a poly-G polypeptide chain was performed to assess the intrinsic ability of the (jigsaw puzzle) surface fit of side-chains in generating a viable core, without any reference to non-core residues. Interestingly, even on a poly-G chain the densely packed cluster centered about 59 (consisting of sites 120, 122, 134, 151, 47, 45, 29) were predicted almost identical to the native sequence (181_R; Table 3), with the exceptions of sites 122,29 solved as F and I, respectively (instead of M and V in native). However, post 31, F was predicted at 181 (<SN^{SC}>: 292.2) to the exclusion of phenylalanine at 76 (288.5); which was followed by markedly different residue substitutions (with respect to native) from sites 179 to 161. The initial raw core obtained on a poly-G polypeptide chain had an overall <SN^{SC}> of 234.2. In Level 2, the convergence procedure was performed twice - first for the raw core on a poly-G polypeptide chain followed by another set of iterations upon the inclusion of the non-core native sequence. The final <SN^{SC}> of the converged core on a native non-core sequence was 387.1. Convergence followed by optimization (identical in procedure described above) yielded 181_O (Table 3) which differed from ALL_O only in four sites (76 181 179 and 164), all pertaining to the loosely packed region of the core. The most notable difference was for sites 76 and 181, where it appeared that the substitution of phenylalanine at either site was mutually exclusive. Optimization of 181_C (Table 5) led to the transitions 120F → L, 122F → M, 79 L → V, 161F → V (181_O), and similar to the situation in Level 1, M was initially replaced by L at site 122, due to its failure in satisfying additional constraints in surface complementarity and overlap. In general, a gradual drift toward increased core volumes was also observed during the convergence process (181_R: 3391.6; 181_C: 3409.3; 181_O: 3286.3 Å^3). Both the predicted core sequences (ALL_O and 181_O) were identical from sites 59 to 31, with significant heterogeneity in sequence thereafter (Table 3).

3.2 | Stability of converged and optimized cores: MD simulation study

In order to estimate relative thermal stabilities, unfolding simulations of the native protein and designed sequences were performed at temperatures 310, 385, and 410 K (see Section 2). The backbone RMSD's (root mean square deviations) with respect to the native crystal structure indicated complete unfolding of all sequences by 410 K, while at 310K the designed sequences were for the most part "native-like"—implying the conservation of the cyclophilin fold and regular secondary structural elements, throughout the duration of the simulation. Therefore, 385 K was deemed the optimal temperature where there was maximum discrimination in parameters indicative of thermal stability (Figure S3 and Figure 4). Overall, the backbone RMSD's appeared to indicate relatively enhanced stability of ALL_O, 181_O relative to the converged sequences ALL_C, 181_Cat all three temperatures (Figure S3). A particularly sensitive parameter was the root mean square fluctuation (RMSF) of the residue C_α, which generally indicated greater thermal fluctuations of all the designed sequences with respect to the native protein (Figure 4). Typically, the propensity for greater thermal fluctuation was observed in regions of the polypeptide chain spanning residues 45–75, 100–120, 125–135, and 145–150 at higher temperatures. Significantly high fluctuation was observed for ALL_C even at 310 K spanning sequence ranges 100–120 (~0.30 nm) and 120–135 (~0.40 nm) different from the native sequence, which was consistently below 0.1 over the entire polypeptide chain. This was somewhat in contrast to 181_C and 181_O whose RMSF values were in fairly close agreement with the native sequence at the

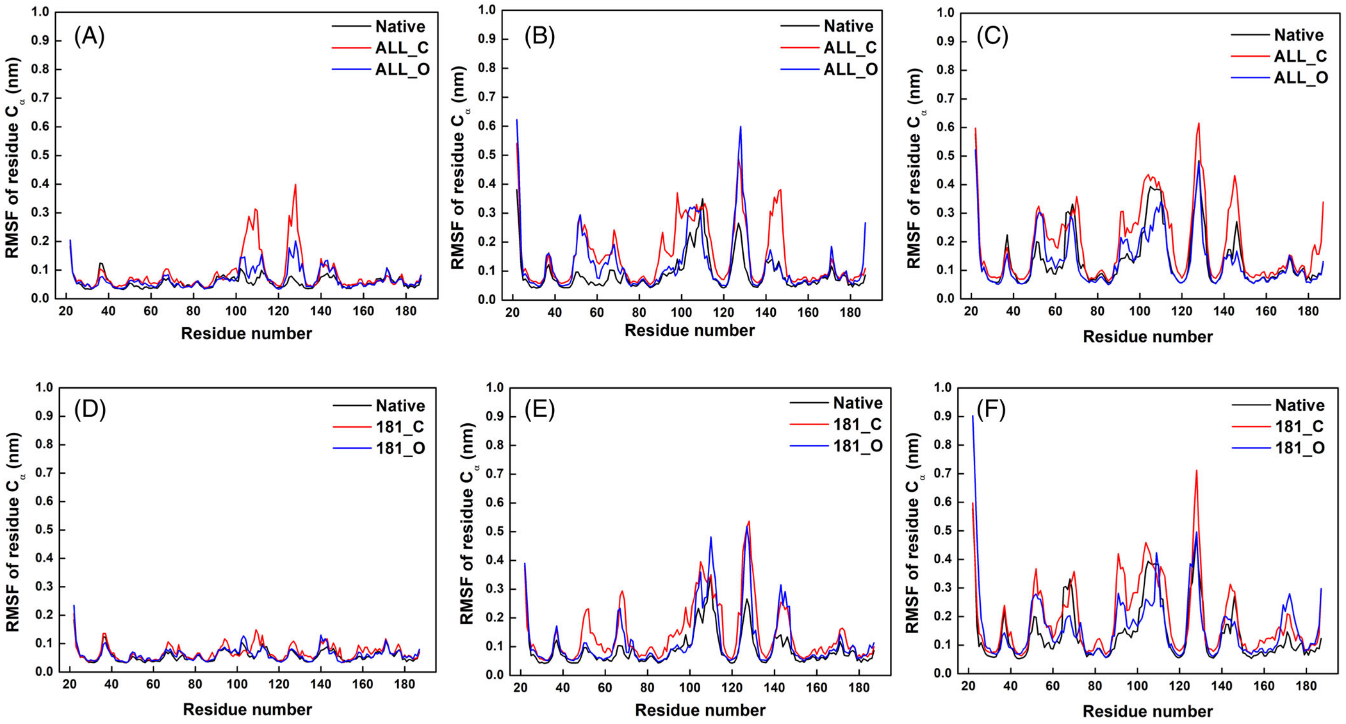


FIGURE 4 Comparison of stabilities of converged (ALL_C, 181_C) and optimized sequences (ALL_O, 181_O) from MD simulation studies. The root-mean-square fluctuations (RMSF) of C_α atoms have been plotted as a function of residue number in case of ALL sequences at three different temperatures, namely, (A) 310, (B) 385, and (C) 410 K; the same parameter has also been plotted for 181 sequences at three different temperatures, (D) 310, (E) 385, and (F) 410 K. The color coding for the sequences have been provided at the upper right-hand corner of the figure

same temperature. However, at 385 K, an appreciable rise in RMSF for the designed sequences (with respect to native) was observed spanning polypeptide segments 45–65 and 120–135. The amino acid stretches 52–65 corresponds to helix H1, while the rest are irregular loops interconnecting the regular secondary structural elements of the protein. In the native sequence, the RMSFs for stretches 50–65, 100–120, and 120–135 were approximately 0.1 (or less) and 0.3 nm, respectively (at 385 K), while for almost all the designed sequences, the corresponding values rose to 0.3 and 0.6 nm. Only 181_O exhibited fluctuations comparable to the native sequence around 50–60. Thus, the integrity or the measure of attachment of helix H1 (55–60) to the main body of the protein appeared to be one of the sensitive indicators of protein stability, with all designed sequences exhibiting some measure of thermal destabilization relative to the native protein, more for the converged than the optimized sequences.

3.3 | Experimental characterization of alternative cores

3.3.1 | Spectroscopic study of the optimized core sequences

The experimental characterization of the optimized sequences (ALL_O, 181_O) was performed by measuring circular dichroism (CD) to confirm whether the cyclophilin fold or in any case the native proportion of regular secondary structural elements in the designed protein had been attained. This was followed by denaturant (guanidium hydrochloride) induced unfolding experiments using tryptophan fluorescence. The LdCyp protein consists of a sole tryptophan (W143) situated onto the small helix H2. In general, tryptophan fluorescence is highly sensitive to its local microenvironment and its use in monitoring protein folding and stability is well-established.[71,72] Here, it has been used to determine the relative stability of the designed sequences with respect to native LdCyp. The computations coupled to the MD simulations appeared to indicate that the optimized sequences (ALL_O and 181_O) would probably have the highest probability of attaining the design target. Belying all expectations, the CD traces of both sequences exhibited extreme deviations from the native spectra (Figure S4a) and resembled the spectra corresponding to the unfolded state of native LdCyp (Figure S4b). Also, the high degree structural perturbations of both sequences were indicated by their extremely low expressive yields (Figure S4c). Thus, a radically different pattern in fluorescence intensity and emission maxima changes upon GdmCl addition (with respect to native) (Figure S4d,e), the inability to obtain a regular unfolding curve (fluorescence intensity versus denaturant concentration: Figure S4f), definitely indicated extreme deviations from the native cyclophilin fold and lack of success in achieving design objectives.

3.3.2 | Characterization of the single mutants

On account of these negative experimental results, it was decided to rebuild the designed cores by systematically mutating one core site at a time beginning with the native protein (Table 6) to identify unviable residue combinations. The mutations common to both ALL_O and 181_O (with respect to the native sequence) are V29I, A52 V, M122 L, and I45 V. Among these, the first three mutations were studied individually, along with I164 M which was present exclusively in ALL_O (Figure 5A,B).

All the four single mutations V29I, A52V, M122L, and I164M on LdCyp exhibited similar characteristics in CD spectra [Figure 5C) as that of the native, including two prominent negative peaks, around 208 and 225 nm. These observations suggested the conservation of the overall fold in the case of all these four mutants, despite the fact that partial destabilization (relative to native) was observed in denaturant-induced unfolding studies. The fluorescence spectra of the native protein as well as all the mutants exhibited emission maxima around 343 nm in the absence of denaturant and showed a significant red-shift up to 350 nm at the highest GdmCl concentration. The denaturation curves (of tryptophan fluorescence intensities; Figure 5D versus denaturant concentration) were optimally fitted to a three-state unfolding equation which exhibited two free energies of denaturation, namely ΔG_{NI} (H_2O) and ΔG_{ND} (H_2O) and two transition mid-points [namely C_m(N → D) and C_m(N → I)]. The stability of the mutants was compared to the native based on these thermodynamic parameters (Table 7).

The mid-point (Table 7) for the major transition [C_m(N → D)] in case of the four mutants V29I, A52 V, I164 M, and M122 L were 1.23, 1.17, 1.11, and 1.05 M, respectively, somewhat lesser than the corresponding native value (1.59 M). In addition, the mid-point for the minor transition [C_m(N → I)] for all the mutants also recorded a decrease (Table 7). A corresponding decline in ΔG's was observed (ΔG_{ND} (H_2O)/ΔG_{NI} (H_2O): V29I-7.59/4.76; A52V-5.31/3.82; I164M-8.30/3.91; M122L-5.76/3.42 kcal/mol) indicating partial destabilization of all the mutants relative to the native structure (16.93/5.01). Among the mutants, V29I, A52V on one hand and I164M, M122L on the other, exhibited comparable ΔG and C_m values, which is reflected in the corresponding denaturation curves (Figure 5D).

3.3.3 | Sequential reconstruction of ALL_O and experimental validation of the mutants

In the sequential reconstruction of the designed chimerical proteins, it is perhaps best to follow a path to the required design target. From the single mutation data, it appeared that A52V and V29I involved the least perturbation to the native fold and thus the double mutant V29I + A52V (ALL_O2′) was initially characterized (Figure 6A,B), demonstrating complete conservation of the native fold (CD data)

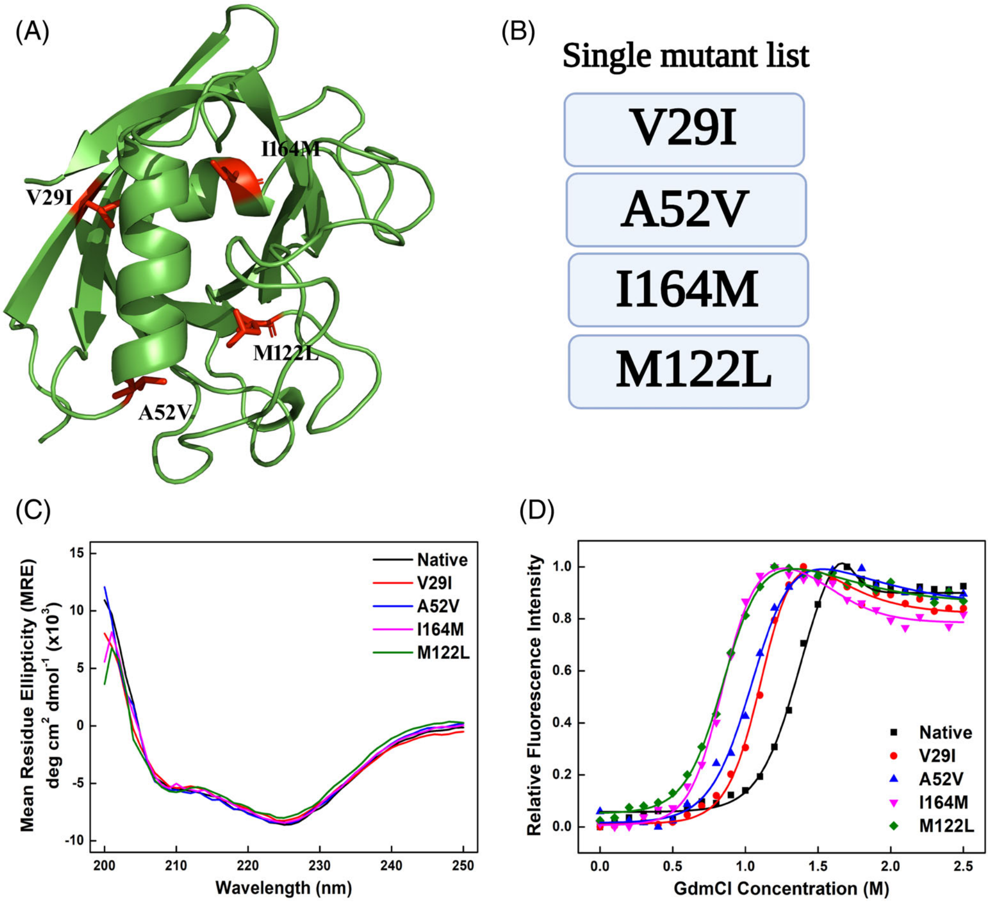
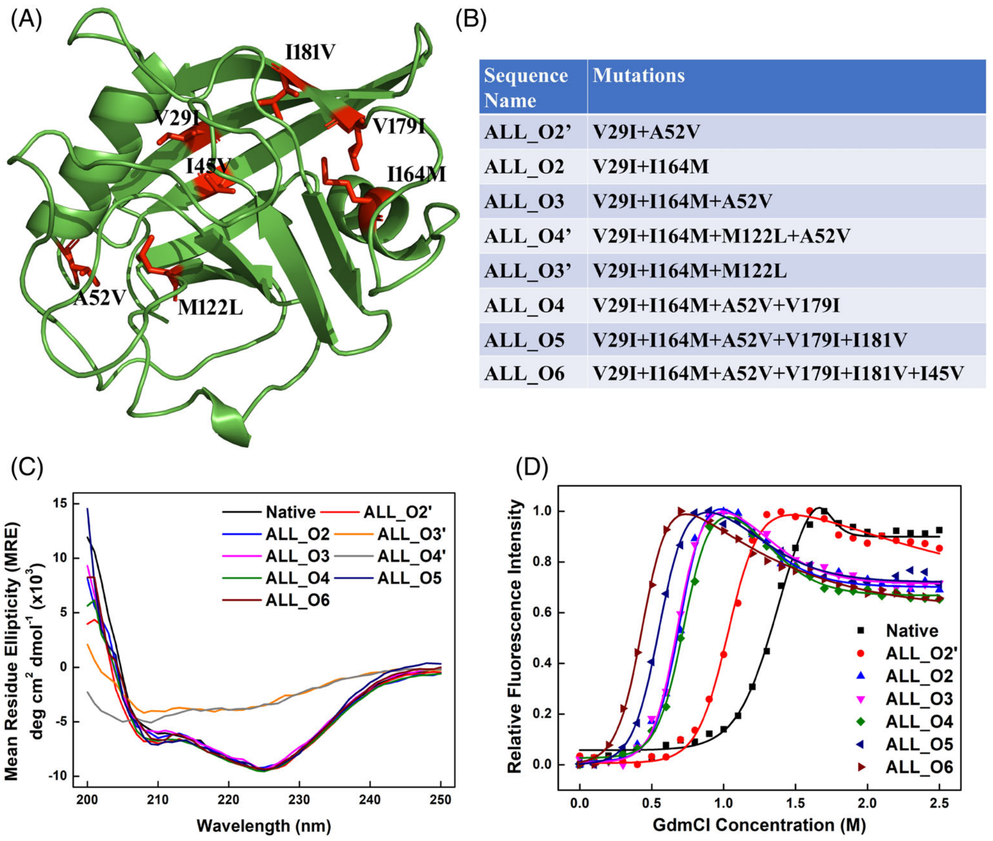


FIGURE 5 Experimental characterization of single mutant proteins. (A) Four mutations, namely, A52 V, V29I, M122 L, and I164 M have been shown in the LdCyp structure, (B) list of mutations that have been characterized experimentally, (C) circular dichroism (CD) spectra of the mutant proteins, (D) denaturant GdmCl induced unfolding of the mutant proteins monitored by intrinsic tryptophan fluorescence where the change in fluorescence intensity have been plotted as a function of denaturant concentration. The color coding of the respective sequences has been provided at the upper right-hand corner of the panel (C) and at the lower right-hand corner for panel (D)

TABLE 7 Thermodynamic parameters of the single mutants and core variants. The parameters have been obtained from the fitting of the unfolding curves to a three-state unfolding equation and the unfolding curves were obtained by measuring the intrinsic tryptophan fluorescence during the GdmCl-mediate unfolding

Protein	$\Delta G_{ND}(H_2O)$ (kcal/mol)	m_{ND} (kcal/ mol M^{-1})	C_m (N → D) (M)	ΔG_{NI} (H_2O) (kcal/mol)	m_{NI} (kcal/ mol M^{-1})	C_m(N → I) (M)	Adjusted R^2
Native	16.93 ± 2.64	10.63 ± 1.57	1.59	5.01 ± 0.08	3.62 ± 0.06	1.38	1.00
V29I	7.59 ± 1.04	6.14 ± 0.71	1.23	4.76 ± 0.52	4.10 ± 0.45	1.16	1.00
A52V	5.31 ± 0.66	4.51 ± 0.47	1.17	3.82 ± 0.51	3.38 ± 0.45	1.13	0.99
I164M	8.30 ± 1.79	7.44 ± 1.18	1.11	3.91 ± 0.37	4.63 ± 0.44	0.84	1.00
M122L	5.76 ± 0.78	5.48 ± 0.52	1.05	3.42 ± 0.27	3.94 ± 0.31	0.86	1.00
ALL_O2′	7.13 ± 0.88	5.90 ± 0.68	1.20	5.11 ± 0.61	4.86 ± 0.59	1.05	0.99
ALL_O2	7.19 ± 0.90	8.42 ± 0.91	0.85	4.12 ± 0.46	5.82 ± 0.66	0.71	0.99
ALL_O3	7.44 ± 0.85	8.64 ± 0.83	0.86	4.14 ± 0.37	6.05 ± 0.56	0.68	1.00
ALL_O4	7.80 ± 0.81	8.54 ± 0.72	0.91	4.15 ± 0.31	5.68 ± 0.42	0.73	1.00
ALL_O5	5.89 ± 0.62	7.96 ± 0.67	0.73	3.19 ± 0.27	5.72 ± 0.49	0.58	1.00
ALL_O6	3.36 ± 0.17	6.58 ± 0.37	0.51	2.81 ± 0.15	5.63 ± 0.34	0.50	1.00
181_O2	6.36 ± 1.95	4.43 ± 1.09	1.43	2.47 ± 0.17	2.01 ± 0.15	1.22	1.00
181_O3	6.05 ± 0.98	5.73 ± 0.67	1.05	2.13 ± 0.16	2.80 ± 0.22	0.76	0.99
181_O4	3.18 ± 0.21	3.87 ± 0.20	0.82	1.64 ± 0.11	2.74 ± 0.20	0.59	0.99
181_O5	3.54 ± 0.22	6.11 ± 0.40	0.57	2.50 ± 0.18	4.99 ± 0.38	0.50	1.00

Note: ΔG_{ND} (H_2O) and ΔG_{NI} (H_2O) are the free-energies of unfolding in the absence of any denaturant (Equations (5) and (6), while m_{ND} and m_{NI} are the slopes of denaturation process N↔D and N↔I respectively. The transition mid-points [C_m(N↔D) and C_m(N↔I)] are the denaturant concentration for which ΔG_{ND} and ΔG_{NI} are zero.



Sequence Name	Mutations
ALL_O2'	V29I+A52V
ALL_O2	V29I+I164M
ALL_O3	V29I+I164M+A52V
ALL_O4'	V29I+I164M+M122L+A52V
ALL_O3'	V29I+I164M+M122L
ALL_O4	V29I+I164M+A52V+V179I
ALL_O5	V29I+I164M+A52V+V179I+I181V
ALL_O6	V29I+I164M+A52V+V179I+I181V+I45V

FIGURE 6 Experimental characterization of the sequences ALL_O2′, ALL_O2, ALL_O3, ALL_O3′, ALL_O4, ALL_O4′, ALL_O5 and ALL_O6 using spectroscopic methods, (A) the position of the mutations with respect to LdCyp crystal structure has been shown, (B) list of all the mutations present in different sequences have been listed, (C) the CD spectra of all the mutant proteins (D) GdmCl-mediated unfolding characterization of the mutant proteins monitored by tryptophan fluorescence (The unfolding spectra of ALL_O3′ and ALL_O4′ have been excluded due to the inability to obtain a proper unfolding curve). Color coding of the sequences has been given in the upper right corner of the panel (C) and the lower right corner for panel (D)

(Figure 6C) and stability comparable to the individual mutants V29I and A52V (Table 7). Another double mutant ALL_O2 (V29I + I164M) was also characterized, followed by the inclusion of I164M in ALL_O2′ resulting in ALL_O3(A52V + I164M + V29I). Although the ΔG values of ALL_O2 and ALL_O3 were not significantly different from double mutant ALL_O2′ (ALL_O2′-7.13, ALL_O2-7.19, ALL_O3-7.44 kcal/mol) there was a drop in C_m values (ALL_O2′-1.20; ALL_O2-0.85; ALL_O3-0.86 M) (Table 7, Figure 6D). The native fold, however, appeared to be fully conserved in these chimerical proteins, borne out by the superposition of their respective CD spectra (Figure 6C).

With the addition of M122L upon ALL_O3 (resulting in ALL_O4′: A52V+I164M+V29I+M122L) (Figure 6A,B) there was again evidence for severe structural distortions similar to the experimental situation in ALL_O and 181_O. Even here (ALL_O4′) there appeared extreme deviations in CD spectra (Figure 6C) coupled to a significant decrease in protein yield and a highly altered denaturation pattern (Data not shown). It thus became obvious that the configuration of residues (A52V+I164M+V29I+M122L) was unviable in the context of the cyclophilin fold. Retracing one step, the experiments were then repeated on the combination (ALL_O3′: A52V + I164M + M122L) with results almost identical to ALL_O4′ (Figure 6C). It will be recalled that in the computational optimization step leucine was selected in favor of methionine at site 122, due to its failure in satisfying additional constraints in surface complementarity and overlap. Therefore, at this stage it was decided to relax the computational constraints at site 122, thereby reverting to methionine (as in native) and restart the core reconstruction process from ALL_O3 (A52V+I164 M+V29I).

Thereafter, three successive constructs were examined by single-step mutations starting from ALL_O3, leading to ALL_O4 (V29I + I164M + A52V + V179I; V179I upon ALL_O3), ALL_O5 (V29I + I164M + A52V + V179I + I181V; I181V upon ALL_O4) and ALL_O6 (V29I + I164M + A52V + V179I + I181V + I45V; I45V upon ALL_O5). No change in the secondary structural content and three-dimensional fold (as evident from CD spectra in Figure 6C), protein yield, and unfolding pattern were observed for these three constructs. It thus appeared that the mutation M122L was primarily responsible for fold collapse in ALL_O, or at least its incompatibility with the set of six mutations in ALL_O6 (as ALL_O can be derived by a single mutation M122L on ALL_O6). The unfolding curves from ALL_O2 to ALL_O4 were almost superimposable while a significant left-shift was observed for ALL_O5 and ALL_O6, leading to lower transition midpoints (Figure 6D). There was also a progressive corresponding decline in ΔG values due to mutations subsequent to ALL_O4 (from 7.80 to 3.36 kcal/mol: Table 7). The ΔG and C_m values of minor transition [ΔG_{NI} (H_2O) and m_{NI}] also followed a similar trend (Table 7). Thus, the mutations primarily responsible for the decline in stability (Figure 6D) despite conserving the cyclophilin fold, appear to accrue from I164M (on ALL_O2), I181V (on ALL_O4), and I45V (on ALL_O5). Therefore, to sum up, with the exception of M122L, the series of 6 mutations in LdCyp conserve the basic cyclophilin fold, despite a progressive decline in stability. ALL_O3′ and ALL_O4′ were particularly destabilized even at zero denaturant concentration while all other mutants appeared to retain the folded state in the absence of denaturant (Figure S5). But for the replacement M122L (as a consequence of cutoffs on methionine), the design objective in terms of native (cyclophilin) fold acquisition would have been achieved in ALL_O.

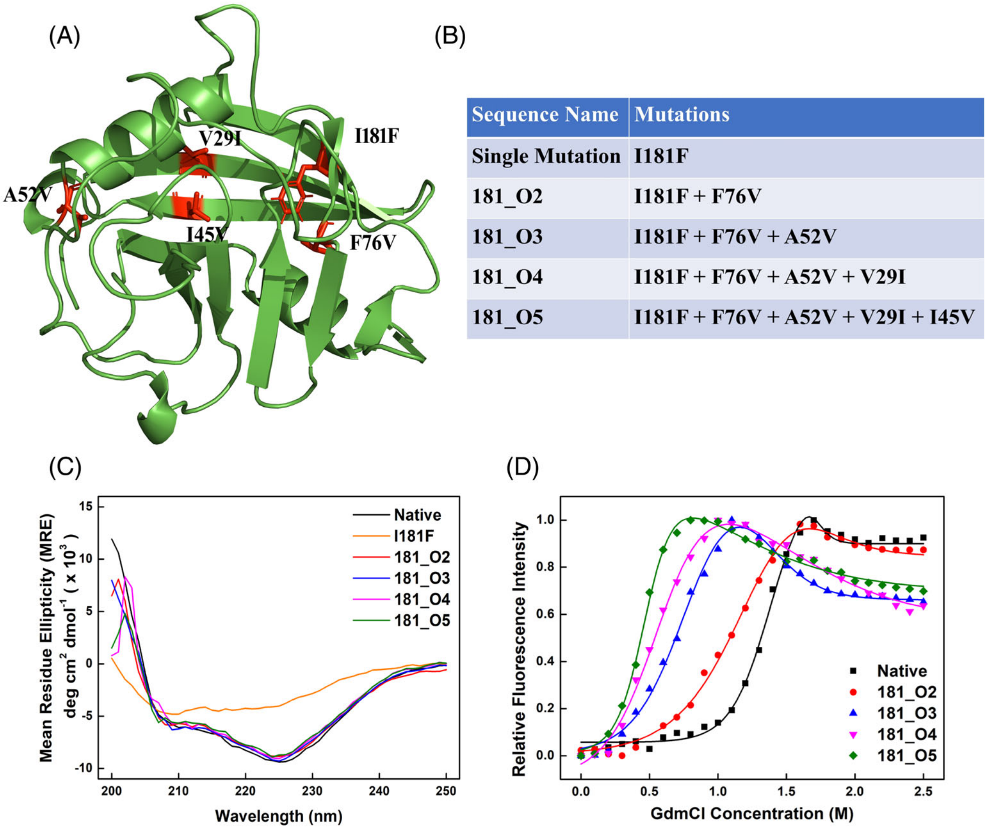
3.3.4 | Reconstruction of 181_O and their experimental validation

The mutations common to ALL_O and 181_O are V29I, A52V, M122L, and I45V, all single mutants which had previously been characterized (with the exception of I45V). As the structural consequences of the two coupled mutations F76V and I181F were unknown, it was decided to initiate the rebuilding of 181_O, beginning with the mutations I181F followed byF76V (Figure 7A,B). At the very first step, it became obvious that the two bulky residues (F) simultaneously present at 76 and 181, led to the unraveling of the protein fold (Figure 7C), as all experimental parameters resembled those of 181_O and ALL_O. However, the compensatory mutation in 181_O2 (I181F +F76V) evidently led to the restoration of the native LdCyp fold (as per CD traces (Figure 7C) with fairly native-like C_m values (C_m (N → D) = 1.43 M and C_m (N → I) = 1.22 M), though with the reduction in ΔG's [ΔG_{ND} (H_2O)-6.36 and ΔG_{NI} (H_2O)-2.13 kcal/mol] (Table 7).

Additional mutations A52V upon 181_O2 resulting in 181_O3 (I181F + F76V + A52V) followed by 181_O4 (I181F + F76V + A52V + V29I; V29I mutation upon 181_O3) and 181_O5 (I181F + F76V + A52V + V29I + I45V; I45V mutation upon 181_O4), all led to almost identical CD traces, which were superimposable on the native LdCyp trace. Between 181_O2(ΔG_{ND} (H_2O)-6.36, ΔG_{NI} (H_2O)-2.47 kcal/mol) to181_O3 (ΔG_{ND} (H_2O)-6.05, ΔG_{NI} (H_2O)-2.13 kcal/mol) there was only a marginal difference in ΔG's, with a progressive decline in all the thermodynamic parameters (characterizing protein stability) thereafter (Table 7, Figure 7D). Notably, the final mutation M122L to 181_O5 leads us back to 181_O which (as mentioned earlier) displays all characteristics of a radically different protein fold compared to native LdCyp. Interestingly, the final design targets 181_O5 and ALL_O6 had remarkably similar unfolding parameters, and the drop in stability subsequent to the mutation I45V, in both cases could possibly be attributed to the weakening of the interaction between residues 45 and 161. Thus, for both ALL_O and 181_O, the single mutation M122L appeared to come in the way of attaining design objectives, and barring this mutation the CD traces were indistinguishable for the native LdCyp trace, and the denaturant induced unfolding curves were of a similar nature.

4 | DISCUSSION

Most globular proteins are marginally stable (ΔG ~ 10–15 kcal/mol) and it has been hypothesized that evolutionary constraints could be far more stringent for protein function rather than thermal stability.[73] In other words, threshold thermodynamic stability could be acquired



Sequence Name	Mutations
Single Mutation	I181F
181_O2	I181F + F76V
181_O3	I181F + F76V + A52V
181_O4	I181F + F76V + A52V + V29I
181_O5	I181F + F76V + A52V + V29I + I45V

FIGURE 7 Experimental characterization of the mutant I181F and sequences 181_O2, 181_O3, 181_O4, and 181_O5 using spectroscopic methods, (A) the position of the mutations with respect to LdCyp crystal structure has been shown, (b) list of all the mutations present in different sequences have been listed, (C) the CD spectra of all the mutant proteins (D) GdmCl-mediated unfolding characterization of the mutant proteins monitored by tryptophan fluorescence (The unfolding spectra of I181F has been excluded due to the inability to obtain a proper unfolding curve). Color coding of the sequences has been given in the upper right corner of the panel (C) and the lower right corner for panel (D)

by a protein in the course of genetic drift, with positive natural selection actively optimizing for catalytic function thereafter.[74] There could also be a trade-off between functionality and stability so as to allow greater conformational flexibility to the protein molecule, for more efficient function.[75] This principle is probably consistent with the fact that alternative cores conserving native fold are indeed possible, which in several instances have been shown to have enhanced thermodynamic stability (with respect to native).[43] In this case, however, it is highly likely that the native cyclophilin core sequence (in LdCyp) is already at an advanced stage of optimization and therefore most modifications to the core will in all probability be destabilizing. One successful approach to core or protein design is via the "consensus sequence" where it is assumed that the more frequent occurrence of a residue at a specific site, the more its contribution to fold stabilization. Extraction of the consensus sequence for the cyclophilin fold (from a limited database of homologous sequences spanning an extensive range of sequence identities) was almost identical to the LdCyp sequence (with the sole exception at the 52nd site) (Table S3).

One of the primary motivations behind this work is to test the efficacy of the jigsaw puzzle model in predicting the hydrophobic core. To this end, the selection of core residues has been based on a surface complementarity function, in order to bias the solution in terms of the optimal fit of side-chain surfaces (of core residues), rather in the manner of assembling a three-dimensional jigsaw puzzle. Probably as expected (from previous calculations performed in the laboratory[54]) only a subset of residues exhibit high specificity contacts resembling the pieces of a jigsaw puzzle and in this particular case involved residues in the neighborhood or linked to F59. The highly specific side-chain contact between F59 and F151 served as a benchmark in these calculations as it appeared to be highly significant both in terms of the surface complementarity measure ($<SN^{SC}>$) and also structurally, enabling anchoring of helix H1 onto the eight stranded β-barrel. Notably, F59 was the highest degree node (Figure 2) and thus, the network (generated from the main chain coordinates alone) representation of the core could prove to be an indicator, where we would expect such high specificity contacts to occur. Generally, such high specificity contacts could be expected to involve nodes of a very high degree, and F59 and its associated links appear to be under stringent packing constraints. Alternative packing arrangements can only be expected in regions where there is a relaxation in packing constraints reflected in nodes with a somewhat sparse distribution of links (V79, V179, V161, I164, and I85). These observations appear to be borne out by the distribution and pattern of substitutions in the alternative cores.

The prediction algorithm applied in this work appears to demonstrate that there is significant information intrinsic to the core without any reference to non-core residues, as evident from the fairly significant sequence identity of the predicted cores, even on a poly-G polypeptide chain, with respect to native (Table 3). It could be likely that subsequent to hydrophobic collapse in a nascent folding polypeptide chain, highly specific (jigsaw puzzle-like) side-chain contacts within the core are essential to the acquisition of the native and viable internal architecture of the protein.

The network representation of the core could also provide useful information on the sequence of atomic events leading to the folded protein. Equilibrium unfolding by heat and chemical denaturant (GdmCl) indicated a non-two state unfolding, with one intermediate.[70] Kinetic unfolding experiments on (the LdCyp homolog) cyclophilin from *E. Coli* showed an extraordinary rapid (hydrophobic collapse) burst phase (with a rate constant of 700 s^{-1}), resulting in the fully folded native protein.[76] In fact, this was the first instance of a protein greater than 100 amino acids in length (164 residues) with such a rapid rate of folding. Detailed unfolding simulations of LdCyp (by MD simulations utilizing metrics based on surface complementarity and overlap) showed that the dissociation of the specific link F59-F151 and general melting of the edges connected to the network hub (F59), was immediately succeeded or in tandem with major transitions, resulting in the unraveling of the structure.[77] Given the rapid folding of this class of proteins, it appears highly likely that the (jigsaw puzzle-like) specific links associated with the hub (F59) coalesce very early in the folding process.

The calculation confirmed the imperative need for a “design cycle” in either repacking or predicting the hydrophobic core, as there could exist highly sensitive strategic “hot spots” or perhaps “tipping points” which could exercise an inordinate influence in the constitution of a viable core. In the present study, M122 was one such site, where in LdCyp the substitution of leucine (at 122) led to unraveling of the (native) cyclophilin fold in both ALL_O and 181_O. This appears all the more surprising as both methionine and leucine have similar volumes and this could be considered a classical example where the change in shape (and therefore specificity) at a single site could have such far-reaching consequences on the overall structure. Repeated failure to achieve the design target in ALL_O was finally traced to this residue which was found to be consistently incompatible with other mutations (Figure 6). This is somewhat reminiscent of metamorphic proteins where a few selective mutations can lead to fold transition, though in this case most probably replacement of methionine with leucine (at 122) led to protein unfolding rather than alternative fold acquisition.

Generally, the replacement of a buried methionine by an equally buried leucine could be expected to be a stabilizing mutation due to a favorable solvent transfer term coupled to reduced entropic penalty (for leucine upon burial).[78] Yet, experiments in T4 lysozyme confirmed that the methionine → leucine could be highly context-dependent, whereby in two instances (M6L and M102L) the mutation was found to be actually destabilizing.[78] Conversely, in a remarkable experiment, the replacement of 10 adjacent residues by methionine in the core of T4 lysozyme still conserved fold and produced an active protein, though with progressively declining thermal stabilities.[12] Methionine has been regularly used in the repertoire of hydrophobic residues in the design of protein cores, though its replacement could be fraught with unforeseen consequences.[45,79] The residue position corresponding to the site 122 is highly conserved for methionine in the cyclophilin fold, though interestingly two sequences 3BKP and 2OSE have valine and leucine at the same site. Both these sequences have numerous other substitutions with respect to the consensus sequence at critical core sites (31, 33, 45, 48, 52, 55, 63, 85, 120, 151, 161, 164, and 179; see Table S3), while 2OSE has been shown to lack PPIase catalytic activity in addition to being the only known aggregated multimeric cyclophilin till date.[80]

In an attempt to understand the structural consequences of the mutation M122L, a MD simulation was performed on the constructs ALL_O3 (V29I + A52V + I164M) and ALL_O4′ (V29I + A52V + I164M + M122L), as fold disruption was first noted in the transition between these two constructs. Although the fluctuations of ALL_O4′ did not show significant alterations from that of ALL_O3 at 310 K (Figure 8A) higher main-chain fluctuations (in terms of RMSF at 385 K) was observed for ALL_O4′ (relative to ALL_O3 and the native sequence) involving residues 60–80 (in the neighborhood of F59, spanning helix H1) and 82–110 (strand S4 and connecting loops) (Figure 8B). Again at 410 K, both the ALL_O3 and ALL_O4′ exhibited a higher degree of fluctuations, which were not much different from each other, as well as from the native LdCyp (Figure 8C). Concurrently, there was also a drop in the number of side-chain contacts of F59 for ALL_O4′ (12.50) compared to ALL_O3 (18.88) and the native sequence (17.15), at the same temperature (Figure 8D). Thus, qualitatively it appeared that the neighborhood of the hub F59 could be most sensitive to mutations, altering the efficient packing of helix H1 to the β-barrel, which could have untold consequences for overall protein fold stabilization.

One of the drawbacks of the predictive algorithm lies in its tendency to gradually drift toward cores of elevated volumes, though core optimization procedures did alleviate this problem to a limited extent (Table 3). Previous design experience indicated that cores tend to over pack on account of softening/damping the Lennard-Jones repulsive potential energy term and the situation could be remedied by steepening the penalty on atomic overlaps.[81] This work attempts to estimate the information intrinsic to the core (based on a minimalist single term function) and it appears highly probable that even though the surface complementarity measure ($<SN^{SC}>$) drops on the atomic overlap, the absence of a (much more) stringent repulsive term in the function, could lead to inflation in core volume. Since the jigsaw puzzle model is concerned primarily with the packing of side-chain atoms within a protein, the final selection of designed sequences entirely dispenses with the interaction of the side chain to main-chain atoms. Although the fixed main chain coordinates (used in this work) were from the high-resolution crystal structure of LdCyp, there are striking examples in the literature where molten cores in de novo designed structures coalesced into well-packed cores upon the formally including flexible backbone and refinement of main chain coordinates, in the design process.[39,44] Generally, backbone optimization leads to improvements in the thermal stability of the designed protein[82] and dramatic thermal stabilization of completely novel proteins has been obtained upon introducing backbone flexibility.[41] What is unambiguous is that the predicted core sequences are sensitive to even limited (~2.0 Å) relaxation of backbone coordinates.[82] Although, the current calculation assumes a fixed backbone, ignoring main chain relaxation upon multiple mutations could be considered to be a lacuna in the method, as even relatively limited main chain readjustments (as a

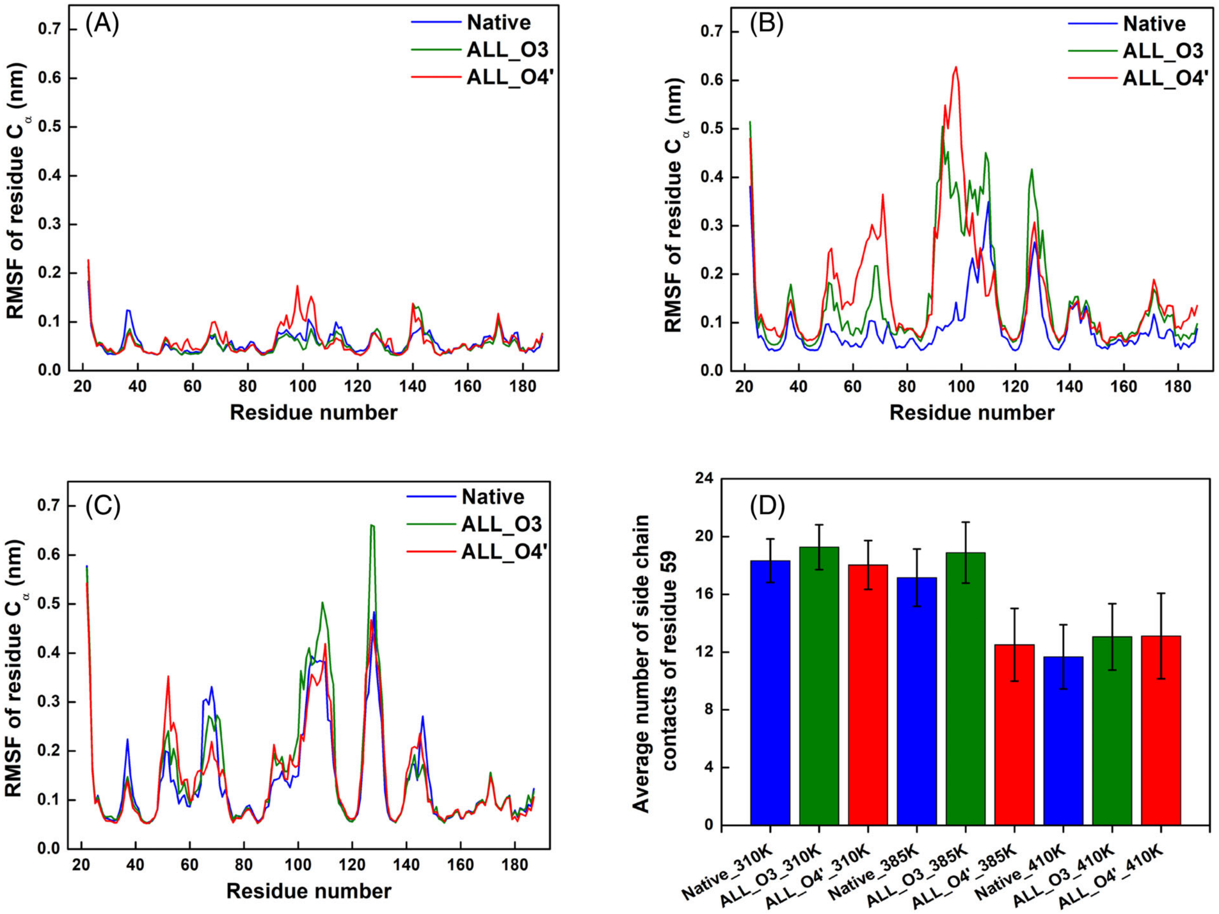


FIGURE 8 Investigation of the effect of M122L mutation by MD simulation involving ALL_O3 (V29I+A52V+I164M) and ALL_O4′ (V29I + A52V + I164M + M122) (A) root mean square fluctuations (RMSF) of C_α atoms of the two proteins at 310, (B) 385 and (C) 410 K temperature, (D) the side-chain contacts of residue 59 of ALL_O3 and ALL_O4′ during the simulation run at three different temperatures

consequence of mutations) could alter the electrostatic balance of energies in the protein.

As mentioned previously, the cyclophilin core naturally exists as a highly optimized ensemble, as the addition of mutants to the native core (discounting M122) invariably leads to progressive thermal destabilization of the altered protein. In order to provide some measure of justification to the increasingly negative (experimental) $\Delta\Delta G_{ND}$ values (Table 7), three programs (SDM2,[83] Dynamut 2,[84] and INPS-3D[85]) were adapted to obtain theoretical ΔΔG estimates to compare with the experiment (Table 8). Although all the three programs were generally successful in predicting the progressive destabilization of LdCyp (with fairly significant correlation with experimental ΔΔG's) several anomalies were observed in the theoretical estimates when compared to experimental values. Of the three programs, INPS-3D was the most successful and among all the computational methods, correctly predicting the relative destabilization of the M122L mutation [experimental −11.17 kcal/mol; theoretical SDM2: 0.29 Dynamut 2: 0.863 and INPS-3D: −1.531 kcal/mol], while SDM2 and Dynamut 2 considered the same mutation to be stabilizing (Table 8). INPS-3D also proved to be the strongest indicator of trouble brewing for the serial mutants M122L + I164M + A52V + V29I, M122L + I164M + V29I (−3.173, −2.838 kcal/mol respectively). With the exception of SDM2, the software also correctly predicted maximum destabilization for ALL_O and 181_O, in their respective series (−6.12 and −6.53 respectively). All the three software however failed in providing a detailed one-to-one correspondence with regard to experimental values. One such lapse is for the pair of mutant structures M122L + I164M + V29I (SDM2: −1.37, Dynamut 2: 0.024, INPS-3D: −2.838 kcal/mol) and I45V + I181V + V179I + I164M + A52V + V29I (−6.53, −4.19, −6.72 kcal/mol respectively), the former being completed unfolded, while the latter preserving fold though with a significant degree of destabilization. It is somewhat problematic to exactly pinpoint the source of differences in the ΔΔG values from these programs as the contributions of the individual "energy" terms are not provided by the software. No correlation was observed between the increase in volume and the experimental ΔΔGs (Correlation coefficient = 0.024, Table 8).

In an attempt to identify the principal causes behind the progressive destabilization of the cyclophilin fold, the Charmm27[53] force field was used to calculate the electrostatic, Lennard-Jones, and geometrical contributions to the overall potential energy of a few mutant structures whose MD simulation were performed (Table S4). By

TABLE 8 Thermodynamic stability analysis of the mutant proteins: Comparison of ΔΔG values derived from the experiment with that of the theoretical values obtained from three web-servers, namely, SDM2, Dynamut 2, and INPS-3D

			Experimental ΔΔG	Theoretical ΔΔG		
			$\Delta\Delta G_{ND}$ (H_2O)	SDM2 server	Dynamut 2 server ΔΔG	INPS-3D ΔΔG
Mutant name	Core volume (Å^3)	Difference in core volume ΔV (Å^3) 3257.8	kcal/mol	Pseudo ΔΔG	kcal/mol	kcal/mol
V29I	3284.5	26.7	−9.34	−0.57	1.004	−0.004
A52V	3309.2	51.4	−11.62	−0.45	0.894	−0.243
M122L	3261.6	3.8	−11.17	0.29	0.863	−1.531
I164M	3254.0	−3.8	−8.63	−0.28	−0.418	−1.354
I181F	3281.0	23.2	–	−1.34	−0.343	−1.927
A52V + V29I	3335.9	78.1	−9.8	−1.02	0.050	−0.331
I164M + V29I	3280.7	22.9	−9.74	−1.20	0.844	−1.402
I164M + A52V + V29I	3332.1	74.3	−9.49	−1.65	−0.230	−1.703
M122L + I164M + A52V + V29I	3335.9	78.1	–	−1.94	−1.100	−3.173
M122L + I164M + V29I	3284.5	26.7	–	−1.37	0.024	−2.838
V179I + I164M + A52V + V29I	3358.8	101.0	−9.13	−2.22	−1.160	−1.820
I181V + V179I + I164M + A52V + V29I	3332.1	74.3	−11.04	−4.26	−2.410	−2.890
I45 V + I181 V + V179I + I164 M + A52 V + V29I	3305.4	47.4	−13.57	−6.41	−3.76	−3.717
M122L + I45V + I181V + V179I + I164M + A52V + V29I	3309.2	51.2	–	−6.12	−4.62	−4.670
F76V + I181F	3231.1	−26.7	−10.57	−3.54	−0.62	−4.182
A52V + F76V + I181F	3282.5	24.7	−10.88	−3.99	−1.19	−4.456
V29I + A52V + F76V + I181F	3309.2	51.4	−13.75	−4.67	−2.02	−4.463
I45V + V29I + A52V + F76V + I181F	3282.5	24.7	−13.39	−6.82	−3.37	−5.250
M122L + I45V + V29I + A52V + F76V + I181F	3286.3	28.5	–	−6.53	−4.19	−6.720

"geometrical" is meant the energies associated with the bond lengths, bond angles, dihedrals, impropers, and the Urey-Bradley component.[53] The difference in the respective energy components (averaged over snapshots) at 310 K (folded) and 410 K (unfolded) simulations were estimated (for the same polypeptide chain) to provide some idea of the difference in enthalpies for the folded and unfolded proteins. In addition, average solvent accessible surface area (SASA) (obtained from in-built GROMACS program) throughout the simulation trajectory were examined for solvent exposure of hydrophobic surfaces (which translates into desolvation energies, Table S4). From this analysis, it appeared that the primary difference in energies (at least with respect to native) appeared to arise from the electrostatic and Lennard-Jones energy terms. The difference in energy for the geometrical term appeared negligible and no exceptional difference in the exposure of hydrophobic surfaces was observed between the mutant structures, taking into account their folded and unfolded states (Table S4).

Thus, in summary, it could tentatively be proposed (pending more detailed calculations) that the progressive loss of stability in the mutant structures probably arises from less favorable electrostatic and van der Waal's interactions in the protein interior as a consequence of amino acid substitutions. It follows that the two major lacunae in the current method lie in the absence of an appropriate repulsive potential in the surface complementary function and inability to assess main chain relaxation of the mutant structures leading to altered electrostatic interactions within the protein.

The fold prediction algorithm adopted in this work biases the selection of residues on the basis of side-chain surface complementarity of core residues, as the initial construction of side chains by SCWRL4.0 involves a spectrum of other energy terms. SCWRL4.0 played a crucial role in this work and has been designed to rapidly determine the side-chain rotamers pertaining to the lowest energy conformation, for a set of residues (given the main chain coordinates). However, it is not optimized for and hence unsuitable in discriminating between two different primary (core) sequences, based upon its intrinsic energy function.

So, finally can the jigsaw puzzle model predict the hydrophobic core of a protein? The answer is a conditional yes-as a "design cycle" involving experiments currently appears unavoidable in achieving design targets. Can the method be used to explore alternative modes of packing in a protein? The answer is an unequivocal yes. Thus, in combination with the full spectrum of energy terms (involving hydrophobicities, solvation terms, etc.) network-based techniques coupled to complementarity principles could provide crucial information in attaining design targets. In addition, electrostatic potential complementarity within proteins could also be explored in situations of ambiguous or loose packing such as those found in amyloids (Discussion S1) or intrinsically disordered proteins. To sum up, this work elucidates both the advantages and the shortcomings of the surface complementarity term (coupled to network-based techniques) in protein design.

ACKNOWLEDGMENT

The authors thank Prof. Dulal Senapati for providing unlimited access to Hitachi F7000 spectrofluorometer at the Chemical Science division, Saha Institute of Nuclear Physics (Kolkata, India). Mr. Jishu Mandal (Indian Institute of Chemical Biology, Kolkata, India) for technical support in circular dichroism spectroscopy experiments. Prof. Krishnananda Chattopadhyay (Indian Institute of Chemical Biology, Kolkata, India), Prof. H. Raghuraman (Saha Institute of Nuclear Physics, Kolkata), and Prof. Sangram Bagh (Saha Institute of Nuclear Physics, Kolkata) for providing valuable suggestions and comments regarding the work. G.B. and S.G. acknowledge SINP, Department of Atomic Energy, Government of India for the award of a Senior Research Fellowship and the financial support from the DBT-RA program in Biotechnology and Life Sciences, respectively.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26321.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Sankar Basu https://orcid.org/0000-0003-1393-1982

REFERENCES

1. Kalinowska B, Banach M, Wiśniowski Z, Konieczny L, Roterman I. Is the hydrophobic core a universal structural element in proteins? *J Mol Model*. 2017;23:31-34.
2. Richards FM. Areas, volumes, packing, and protein structure. *Annu Rev Biophys Bioeng*. 1977;6:151-176.
3. Tóth-Petróczy Á, Tawfik DS. Slow protein evolutionary rates are dictated by surface-core association. *Proc Natl Acad Sci USA*. 2011;108:11151-11156.
4. Amprazi M, Kotsifaki D, Providaki M, et al. Structural plasticity of 4-α-helical bundles exemplified by the puzzle-like molecular assembly of the Rop protein. *Proc Natl Acad Sci USA*. 2014;111:11049-11054.
5. Lim WA, Sauer RT. Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature*. 1989;339:31-36.
6. Eriksson AE, Baase WA, Zhang XJ, et al. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science*. 1992;255:178-183.
7. Lim WA, Farruggio DC, Sauer RT. Structural and energetic consequences of disruptive mutations in a protein Core. *Biochemistry*. 1992;31:4324-4333.
8. Richards FM, Lim WA. An analysis of packing in the protein folding problem. *Q Rev Biophys*. 1994;26:423-498.
9. Levitt M, Gerstein M, Huang E, Subbiah S, Tsai J. Protein folding: the endgame. *Annu Rev Biochem*. 1997;66:549-579.
10. Sandberg WS, Terwilliger TC. Energetics of repacking a protein interior. *Proc Natl Acad Sci U S A*. 1991;88:1706-1710.
11. Lazar GA, Handel TM. Hydrophobic core packing and protein design. *Curr Opin Chem Biol*. 1998;2:675-679.
12. Gassner NC, Baase WA, Matthews BW. A test of the "jigsaw puzzle" model for protein folding by multiple methionine substitutions within the core of T4 lysozyme. *Proc Natl Acad Sci USA*. 1996;93:12155-12158.
13. Banach M, Fabian P, Stapor K, Konieczny L, Roterman I. Structure of the hydrophobic core determines the 3d protein structure-verification by single mutation proteins. *Biomolecules*. 2020;10:767.
14. Hurley JH, Baase WA, Matthews BW. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol*. 1992;224:1143-1159.
15. Eriksson AE, Baase WA, Matthews BW. Similar hydrophobic replacements of Leu99 and Phe153 within the core of T4 lysozyme have different structural and thermodynamic consequences. *J Mol Biol*. 1993;229:747-769.
16. Kauzmann W. Some factors in the interpretation of protein denaturation. *Adv Protein Chem*. 1959;14:1-63.
17. Banach M, Konieczny L, Roterman I. The fuzzy oil drop model, based on hydrophobicity density distribution, generalizes the influence of water environment on protein structure and function. *J Theor Biol*. 2014;359:6-17.
18. Banach M, Prymula K, Jurkowski W, Konieczny L, Roterman I. Fuzzy oil drop model to interpret the structure of antifreeze proteins and their mutants. *J Mol Model*. 2012;18:229-237.
19. Banach M, Roterman I, Prudhomme N, Chomilier J. Hydrophobic core in domains of immunoglobulin-like fold. *J Biomol Structure Dynam*. 2013;32:1583-1600.
20. Behe MJ, Lattman EE, Rose GD. The protein-folding problem: the native fold determines packing, but does packing determine the native fold? *Proc Natl Acad Sci USA*. 1991;88:4195-4199.
21. Kamtekar S, Schiffer JM, Xiong H, Babik JM, Hecht MH. Protein design by binary patterning of polar and nonpolar amino acids. *Science*. 1993;262:1680-1685.
22. Mani SK, Balasubramanian H, Nallusamy S, et al. Sequence and structural analysis of two designed proteins with 88% identity adopting different folds. *Protein Eng Design Select*. 2010;23:911-918.
23. Desjarlais JR, Handel TM. De novo design of the hydrophobic cores of proteins. *Protein Sci*. 1995;4:2006-2018.
24. Malashkevich VN, Higgins CD, Almo SC, Lai JR. A switch from parallel to antiparallel Strand orientation in a coiled-coil X-ray structure via two Core hydrophobic mutations. *Biopolymers*. 2015;104:178-185.
25. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. *Proc Natl Acad Sci USA*. 2009;106:21149-21154.
26. Lazar GA, Desjarlais JR, Handel TM. De novo design of the hydrophobic core of ubiquitin. *Protein Sci*. 1997;6:1167-1178.
27. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. *Proc Natl Acad Sci USA*. 1997;94:10172-10177.
28. Desjarlais JR, Handel TM. Side-chain and backbone flexibility in protein core design. *J Mol Biol*. 1999;290:305-318.
29. Dantas G, Corrent C, Reichow SL, et al. High-resolution structural and thermodynamic analysis of extreme stabilization of human Procarboxypeptidase by computational protein design. *J Mol Biol*. 2007;366:1209-1221.

30. Goldenzweig A, Goldsmith M, Hill SE, et al. Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol Cell*. 2016;63:337-346.
31. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*. 2009;77:778-795.
32. Liang S, Zheng D, Zhang C, Standley DM. Fast and accurate prediction of protein side-chain conformations. *Bioinformatics*. 2011;27:2913-2914.
33. Miao Z, Cao Y, Jiang T. RASP: rapid modeling of protein side chain conformations. *Bioinformatics*. 2011;27:3117-3122.
34. Eyal E, Najmanovich R, Mcconkey BJ, Edelman M, Sobolev V. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J Comput Chem*. 2004;25:712-724.
35. Wernisch L, Hery S, Wodak SJ. Automatic protein design with all atom force-fields by exact and heuristic optimization. *J Mol Biol*. 2000;301:713-736.
36. Harbury PB, Plecs JJ, Tidor B, Alber T, Kim PS. High-resolution protein design with backbone freedom. *Science*. 1998;282:1462-1467.
37. Hill RB, Degrado WF. Solution structure of α2D, a nativelike de novo designed protein. *J Am Chem Soc*. 1998;120:1138-1145.
38. Dahiyat BI, Mayo SL. De novo protein design: fully automated sequence selection. *Science*. 1997;278:82-87.
39. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D. Design of a novel globular protein fold with atomic-level accuracy. *Science*. 2003;302:1364-1368.
40. Liu Y, Kuhlman B. RosettaDesign server for protein design. *Nucleic Acids Res*. 2006;34:235-238.
41. Murphy GS, Mills JL, Miley MJ, Machius M, Szyperski T, Kuhlman B. Increasing sequence diversity with flexible backbone protein design: the complete redesign of a protein hydrophobic Core. *Structure*. 2012;20:1086-1096.
42. Kuhlman B, Baker D. Native protein sequences are close to optimal for their structures. *Proc Natl Acad Sci USA*. 2000;97:10383-10388.
43. Dantas G, Kuhlman B, Callender D, Wong M, Baker D. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J Mol Biol*. 2003;332:449-460.
44. Betz SF, Raleigh DP, DeGrado WF. De novo protein design: from molten globules to native-like states. Current opinion in structural biology. *Curr Opin Struct Biol*. 1993;3:601-610.
45. Lovejoy B, Choe S, Cascio D, McRorie DK, DeGrado WF, Eisenberg D. Crystal structure of a synthetic triple-stranded α-helical bundle. *Science*. 1993;259:1288-1293.
46. Willis MA, Bishop B, Regan L, Brunger AT. Dramatic structural and thermodynamic consequences of repacking a protein's hydrophobic core. *Structure*. 2000;8:1319-1328.
47. Venugopal V, Sen B, Datta AK, et al. Structure of cyclophilin from Leishmania donovani at 1.97 Å resolution. *Acta Crystallogr Sect F: Struct Biol Cryst Commun*. 2007;63:60-64.
48. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. *J Mol Biol*. 1971;55:379-IN4.
49. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982;157:105-132.
50. Eisenberg D, Schwarz E, Komaromy M, Wall R. Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*. 1984;179:125-142.
51. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. *Proteins Struct Funct Genet*. 2000;40:389-408.
52. Ramachandran GN, Sasisekharan V. Conformation of polypeptides and proteins. *Adv Protein Chem*. 1968;23:283-437.
53. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem*. 1983;4:187-217.
54. Banerjee R, Sen M, Bhattacharya D. The jigsaw puzzle model: search for conformational specificity in protein interiors. *J Mol Biol*. 2003;333:211-226.
55. Basu S, Bhattacharyya D, Banerjee R. Mapping the distribution of packing topologies within protein interiors shows predominant preference for specific packing motifs. *BMC Bioinformat*. 2011;12:195.
56. Basu S, Bhattacharyya D, Banerjee R. Self-complementarity within proteins: bridging the gap between binding and folding. *Biophys J*. 2012;102:2605-2614.
57. Basu S, Bhattacharyya D, Banerjee R. Applications of complementarity plot in error detection and structure validation of proteins. *Indian J Biochem Biophys*. 2014;51:188-200.
58. Word JM, Lovell SC, Richardson JS, Richardson DC. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*. 1999;285:1735-1747.
59. Cornell WD, Cieplak P, Bayly CI, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Am J Chem Soc*. 1995;117:5179-5197.
60. Lawrence MC, Colman PM. Shape complementarity at protein/protein interfaces. *J Mol Biol*. 1993;234:946-950.
61. Krivov GG, Shapovalov MV, Dunbrack RL. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*. 2009;77:778-795.
62. Zamyatnin AA. Protein volume in solution. *Prog Biophys Mol Biol*. 1972;24:107-123.
63. Hess B, Kutzner C, Van Der Spoel D, et al. GRGMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput*. 2008;4:435-447.
64. Hess B. P-LINCS: a parallel linear constraint solver for molecular simulation. *J Chem Theory Comput*. 2008;4:116-122.
65. Harvey MJ, De Fabritiis G. An implementation of the smooth particle mesh Ewald method on GPU hardware. *J Chem Theory Comput*. 2009;5:2371-2377.
66. Berendsen HJC, Postma JPM, Van Gunsteren WF, et al. Molecular dynamics with coupling to an external bath. *J Chem Phys*. 1984;81:3684-3690.
67. Parrinello M, Rahman A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys*. 1981;52:7182-7190.
68. Sen B, Venugopal V, Chakraborty A, et al. Amino acid residues of Leishmania donovani cyclophilin key to interaction with its adenosine kinase: biological implications. *Biochemistry*. 2007;46:7832-7843.
69. Venugopal V, Datta AK, Bhattacharyya D, Dasgupta D, Banerjee R. Structure of cyclophilin from Leishmania donovani bound to cyclosporin at 2.6 Å resolution: correlation between structure and thermodynamic data. *Acta Crystallogr D Biol Crystallogr*. 2009;65:1187-1195.
70. Roy S, Basu S, Datta AK, Bhattacharyya D, Banerjee R, Dasgupta D. Equilibrium unfolding of cyclophilin from Leishmania donovani: characterization of intermediate states. *Int J Biol Macromol*. 2014;69:353-360.
71. Biswas G, Ghosh S, Raghuraman H, Banerjee R. Probing conformational transitions of PIN1 from L. major during chemical and thermal denaturation. *Int J Biol Macromol*. 2020;154:904-915.
72. Royer CA. Probing protein folding and conformational transitions with fluorescence. *Chem Rev*. 2006;106:1769-1784.
73. Privalov PL, Khechinashvili NN. A thermodynamic approach to the problem of stabilization of globular protein structure: a calorimetric study. *J Mol Biol*. 1974;86:665-684.
74. Lipman DJ, Wilbur WJ. Modelling neutral and selective evolution of protein folding. *Proc Royal Soc B: Biol Sci*. 1991;245:7-11.
75. Závodszky P, Kardos J, Svingor Á, et al. Adjustment of conformational flexibility is a key event in the thermal adaptation of proteins. *Proc Natl Acad Sci U S A*. 1998;95:7406-7411.
76. Ikura T, Hayano T, Takahashi N, Kuwajima K. Fast folding of Escherichia coli cyclophilin a: a hypothesis of a unique hydrophobic core with a phenylalanine cluster. *J Mol Biol*. 2000;297:791-802.
77. Roy S, Basu S, Dasgupta D, Bhattacharyya D, Banerjee R. The unfolding MD simulations of cyclophilin: analyzed by surface

contact networks and their associated metrics. *PLoS ONE*. 2015;10:1-40.

78. Lipscomb LA, Gassner NC, Snow SD, et al. Context-dependent protein stabilization by methionine-to-leucine substitution shown in T4 lysozyme. *Protein Sci*. 1998;7:765-773.
79. Benítez-Cardoza CG, Stott K, Hirshberg M, Went HM, Woolfson DN, Jackson SE. Exploring sequence/folding space: folding studies on multiple hydrophobic Core mutants of ubiquitin. *Biochemistry*. 2004;43:5195-5203.
80. Thai V, Renesto P, Fowler CA, et al. Structural, biochemical, and in vivo characterization of the first virally encoded Cyclophilin from the Mimivirus. *J Mol Biol*. 2008;378:71-86.
81. Ventura S, Vega MC, Lacroix E, et al. Conformational strain in the hydrophobic core and its implications for protein folding and design. *Nat Struct Biol*. 2002;9:485-493.
82. Yin S, Ding F, Dokholyan NV. Modeling backbone flexibility improves protein stability estimation. *Structure*. 2007;15:1567-1576.
83. Pandurangan AP, Ochoa-Montaño B, Ascher DB, Blundell TL. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res*. 2017;45:W229-W235.
84. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res*. 2018;46:W350-W355.
85. Savojardo C, Fariselli P, Martelli PL, Casadio R. INPS-MD: a web server to predict stability of protein variants from sequence and structure. *Bioinformatics*. 2016;32:2542-2544.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Biswas G, Ghosh S, Basu S, Bhattacharyya D, Datta AK, Banerjee R. Can the jigsaw puzzle model of protein folding re-assemble a hydrophobic core? *Proteins*. 2022;1-23. doi:10.1002/prot.26321