CHAPTER SIX

Network-based strategies for protein characterization

Alessandra Merlotti[a], Giulia Menichetti[b,c], Piero Fariselli[d], Emidio Capriotti[e], and Daniel Remondini[a,*]

[a]Department of Physics and Astronomy, University of Bologna, Bologna, Italy
[b]Center for Complex Network Research, Department of Physics, Northeastern University, Boston, MA, United States
[c]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States
[d]Department of Medical Sciences, University of Torino, Turin, Italy
[e]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy
*Corresponding author: e-mail address: daniel.remondini@unibo.it

Contents

Abstract

Protein structure characterization is fundamental to understand protein properties, such as folding process and protein resistance to thermal stress, up to unveiling organism pathologies (e.g., prion disease). In this chapter, we provide an overview on how the spectral properties of the networks reconstructed from the Protein Contact Map (PCM) can be used to generate informative observables. As a specific case study, we apply two different network approaches to an example protein dataset, for the aim of discriminating protein folding state, and for the reconstruction of protein 3D structure.

1. Introduction

In the last decade several models describing a protein as a network of interacting residues were used for characterizing the relationship between structure and function (Greene, 2012; Grewal & Roy, 2015; Yan et al., 2014).

Advances in Protein Chemistry and Structural Biology, Volume 127
ISSN 1876-1623
https://doi.org/10.1016/bs.apcsb.2021.05.001

The studies of the protein contact network were used for detecting important residues for protein stability and dynamics (Böde et al., 2007; Brinda & Vishveshwara, 2005; Taylor, 2013), active sites (Amitai et al., 2004) and protein folding kinetics (Bagler & Sinha, 2007). In general, the works in the field focus on the properties arising for the local and global topology of the network (Bollobas, 1998). For the topological analysis, the distribution of the degree of nodes and the shortest paths among nodes are considered the main observables for the description of the protein structure. In this work we analyzed the classical network analysis techniques for the study of protein folding. In particular, we focus on the application of the protein contact network analysis to the reconstruction of the protein three-dimensional structure and to the characterization of the protein folding mechanism.

2. The protein folding problem

Protein folding is the process by which the polypeptide chain reaches its native three-dimensional (3D) structure conformation. The Anfinsens' experiments carried out in the 1970s lead to the conclusion that under favorable conditions, protein will fold consistently into its native structure which is encoded in its amino acid sequence (Anfinsen, 1973). Although this view of the folding mechanism has been challenged by new experimental evidence (Dishman & Volkman, 2018), the large amount of crystallographic data collected in the Protein Data Bank (wwPDB consortium, 2019) reinforce the idea of the uniqueness of the folded conformation. The existence of a stable and kinetically accessible native conformation of the proteins determined by the amino acid sequence enhanced the development of several theoretical models and computational methods for studying the protein folding mechanism (Compiani & Capriotti, 2013; Dill, 1990). The majority of the available models and methods focus on three aspects of the same problem related to the prediction of the native structure, the thermodynamics and the kinetics of the folding process (Dill et al., 2007; Dill & MacCallum, 2012). The prediction of the protein structure from the amino acid sequence is a challenging problem that drew the attention of the scientific community at the end of the 1980s when few hundreds protein structures were made available on the Protein Data Bank (Fariselli et al., 2007).

The seminal work from Chothia and Lesk studying the relationship between protein sequence and structure found that homologous proteins retain the same general fold (Chothia & Lesk, 1986). This observation laid the foundation for the development of computational structure prediction methods which rely on detectable similarity with known protein structures or ab initio methods (Baker & Sali, 2001). During the last two decades the Critical Assessment of Structure Protein (CASP) evaluated the quality of the prediction algorithms tracking the progress in the field (Kryshtafovych et al., 2019). Recently, a dramatic improvement of the performance in the prediction of the protein 3D structure was driven by the successful application of deep learning techniques (Senior et al., 2020). Although the prediction of the native conformation of a protein from its sequence achieved an unprecedented level of performance, the folding mechanism description at thermodynamic and kinetic levels is still incomplete. In the last few years statistical and machine learning algorithms have been developed for predicting the stability of a protein structure and the folding rate. Nevertheless, at the current stage reliable and general models for describing the free energy landscape of the folding process are unavailable. In this context, several methods have been developed for predicting protein stability and folding rate (Chang et al., 2015; Magliery, 2015; Sanavia et al., 2020). These approaches rely on protein 3D structure which is used to identify the interacting residues along the amino acid sequence. Such information is essential for estimating the stability of the native conformation and determining the mechanism of the protein folding.

3. Modeling folding kinetics

The study of the folding kinetics is important for calculating the time by which a protein reaches its native conformation and for identifying the formation of metastable conformations during the folding process. Thus, for the characterization of the folding kinetics were developed several theoretical models based on a simplified representation of the protein structure (Compiani & Capriotti, 2013). Depending on the level of cooperativity in the formation of the native conformation, the models of protein folding were classified in three groups: hydrophobic collapse, nucleation–condensation and framework models. The main differences among these

models depend on the role played by the secondary structure in the formation of the native conformation. Among them, the Diffusion-Collision model is one of the first quantitative models for predicting the folding time (Karplus & Weaver, 1976). This hierarchical model represents the protein as a set of partially formed secondary structure elements that reach the native state through stochastic collision events. More recent models based on non-local interactions between residues consider the static 3D of the protein as a proxy for the prediction of the protein folding rate (Gromiha & Selvaraj, 2001; Ivankov & Finkelstein, 2004; Plaxco et al., 1998; Zhou & Zhou, 2002). These methods rely on the observation that the average sequence separation between contacting residues in the native conformation correlates with the folding rate and transition state of single-domain proteins. Thus, the definition of interacting residues assumed an important role in the determination of the folding mechanism, and for the development of more sophisticated methods based on predicted contact maps and machine learning approaches for predicting the folding rate and mechanism (Capriotti & Casadio, 2007; Huang & Gromiha, 2010; Punta & Rost, 2005). In general, all the methods represent the protein as a graph where the nodes are the residues connected by an edge when the distance between two nodes is below a given threshold. Such representation, which is equivalent to a contact map, is used to compute the distribution of non-local interactions among residues.

4. Protein structure representation

In the last two decades the Protein Structure Initiative strongly contributed to the identification of new protein structures (Grabowski et al., 2016). Such information is important for studying the function of a protein that is related to geometrical features defining the secondary structure of the protein and determining its fold (Hrmova & Fincher, 2009). Currently the PDB collects ~177 K structures which are classified in more than 5000 superfamilies and families by the two most popular databases CATH (Sillitoe et al., 2021) and SCOP (Andreeva et al., 2020). Each protein three-dimensional structure is represented by the coordinates of its atoms. For representing a protein composed by n atoms, 3n numbers are needed. An alternative protein structure representation is based on the distance

matrix which is composed by n^2 elements that for the symmetry are reduced to $n(n-1)/2$. Although the distance matrix has more elements than standard representations based on the atom coordinate, it can be advantageous when only low resolution data from NMR are available (Bartoli et al., 2008). A simplified version of the distance matrix is the contact map which considers only the distance between specific atoms of each residue either α or β carbons and a cut-off distance to represent the presence of absence of a contact with a binary number. The possibility of reconstructing the protein structure starting from a reduced representation is an essential aspect for its application to the study of the protein structure. Previous studies have proved that contact maps provide a good representation of the protein backbone (Porto et al., 2004; Vassura et al., 2008). Thus, the contact map, which retains the main information about protein structure, can be used as a proxy for the characterization of the protein folding mechanism.

5. Contact maps and graph Laplacian

Every protein can be represented as a network of interacting particles, where nodes can correspond to single atoms, residues, or even larger motifs, and links to their interactions. Of course, in this approach, a key point consists of how nodes and links are defined. For sake of simplicity, we consider α-Carbons (Cα) as nodes and distance-dependent interactions as links, by imposing that two nodes are connected if their distance d is lower than a specific threshold t. In this way, given the number N of Cα, we could associate to each protein a contact map A, that is a binary NxN symmetric matrix defined as:

$$A_{ij} = \begin{cases} 1 \; if \; d_{ij} \leq t \\ 0 \; if \; d_{ij} > t \end{cases}$$

In network theory, A corresponds to the adjacency matrix of a graph, and represents the starting point for studying the importance of nodes within the network, and the topological structure of the interactions between them.

From A, the Laplacian operator of a network, is derived as:

$$L = D - A$$

where D is the degree matrix, defined as $D_{ij} = k_i \bullet \delta_{ij}$, and k_i represents the degree of node *i*. In particular, the action of L on a N-dimensional lattice corresponds to the discretization of a N-dimensional elastic membrane, where L's eigenvalues represent the frequencies of the normal modes and L's eigenvectors represent the normal mode solutions or eigenfunctions (Biyikoglu et al., 2007). With this analogy in mind, the eigenvalue decomposition of the Laplacian operator corresponds to searching for extremal values of the Rayleigh functional, vectors x that maximize or minimize the mutual distance between nodes in the network, expressed by the following semi-positive quadratic form:

$$\vec{x}^{T} L \vec{x} = \sum_{i \sim j} \left(x_i - x_j\right)^2$$

The trivial solution corresponds to the 0 eigenvalue, in which all nodes have the same spatial coordinates and thus $x_i = x_j$ for every i,j. The non-trivial solutions seek for a minimal distance by imposing the orthogonality with the constant vector. If we hypothesize that the elastic potential schematized by the Laplacian operator is an approximation around the minimum of the Lennard-Jones potential-like function, modeling the interaction between protein residues, the 3D coordinates of Cα can be estimated by the components of the 3 eigenvectors associated with the 3 smallest positive eigenvalues of the Laplacian operator, thus providing a reconstruction of the 3D protein structure up to a linear transformation.

6. Protein folding state discrimination and Laplacian spectrum

In (Menichetti et al., 2016) the properties of the Laplacian spectrum are leveraged to predict protein folding kinetics as two-state, an "all-or-none" type of transition, or as multi-state, in presence of one or more intermediates. The training database for Fisher discriminant analysis consists of 63 manually annotated proteins by Ivankov and Finkelstein (2004)

(25 multi-state, 38 two-state), all proteins with structure available on PDB (https://www.rcsb.org).

PCMs are derived by choosing an upper threshold of 8 Å. Interestingly, among the 5 network observables defined in the paper, whose performances in combination or alone are extremely predictive of folding classes, we find 3 Laplacian-based variables. However, the best accuracy and Matthews correlation coefficient are not achieved by deriving the Laplacian from the original PCM, but by focusing on a modified version that keeps only long-range contacts, while preserving the network connectivity.

With the hypothesis that the most relevant information on folding kinetics is determined by the long-range contacts of the native folded state, a partial removal of the protein backbone, up to the breaking point of the protein into fragments, enhances the role of long-range connections with respect to the protein backbone, while keeping the PCM still connected. The number of diagonals removed varies from protein to protein.

Once the Laplacian spectrum of the modified PCM is computed for each protein, the 3 largest eigenvalues are collected and rescaled by the number of residues N_C. The rescaling corrects for the dependence of the largest eigenvalues λ_N, λ_{N-1}, λ_{N-2} on N_C. According to the vibrational interpretation of the Laplacian, the selected eigenvalues represent the highest vibrational frequencies associated with the small-range structure of the protein, compared to a more global assessment of the long-range vibrations and algebraic connectivity of the protein structure, offered by the Fiedler number (second smallest eigenvalue).

The percentage of correctly classified proteins, when using λ_N, λ_{N-1}, λ_{N-2} separately, is $76.6\% \pm 1.3$, $76.7\% \pm 1.4$, and $77.6\% \pm 1.1$, representing the average values of 10-fold cross-validation over 10,000 resamplings. Similarly, the Matthews correlation coefficient follows as 0.57 ± 0.02, 0.58 ± 0.02, and 0.59 ± 0.02.

Overall, we observe that two-state proteins tend to have larger values of fast-vibrating frequencies, compared to multi-state proteins, and that the vibrational modes (i.e., the corresponding eigenvectors) associated to high frequencies are in general characterized by a strong localization along the vector, corresponding to specific protein regions. If this feature is observed also in our case, and if this can be associated to specific folding/unfolding dynamics, is still an open issue.

7. A case study: Methods for protein 3D-structure reconstruction

In order to compare and to evaluate two different network-based approaches that we choose to use, we observe that a faithful reconstruction of 3D structure, given the "network" information provided by the contact map only, can be considered a good validation that the network framework adopted can characterize properties associated with protein folding. Therefore, in this section we will show two different network-based methods that allow the reconstruction of the 3D coordinates of Cα, starting from their contact maps.

The first method is based on the first three eigenvectors of the Laplacian operator, as explained in Section 5, that exploits the "vibrational" analogy of the Laplacian operator, as describing a set of unit masses connected by springs with equal stiffness, and which first eigenvectors correspond to the largest-scale vibrational modes. The second method was proposed by Lesne et al. (2014), who devised an algorithm called ShRec3D with the aim of reconstructing the 3D structure of chromosomes starting from Hi-C data (Lieberman-Aiden et al., 2009), which allows the mapping of neighboring DNA fragments, generating an output formally equivalent to a protein contact map, despite the physics underlying chromosome and protein 3D configuration is different (Merlotti et al., 2020), since there are no direct chemical bonds between DNA strands, but rather we can talk about a spatial proximity mediated by other factors (like cohesin, histones and CTCF proteins). Moreover, differently from DNA 3D structure which is still largely unknown particularly at a fine scale of the single nucleotides, for our protein dataset we have the ground truth provided by the protein 3D configuration obtained through X-ray crystallography to be compared with our reconstructions.

The ShRec3D algorithm can be divided into two steps: (1) the computation of Cα distances starting from the contact map and (2) the estimation of Cα spatial coordinates starting from their mutual distances. The first step is performed by measuring the distance between two Cα as the length of the shortest path connecting them in the network provided by the contact map. In fact, it is known that the shortest paths s_{ij} between nodes i and j in a symmetric network satisfy the conditions to be considered a metrics: (1) be $= 0$ if $i = j$; (2) be symmetric $s_{ij} = s_{ji}$; (3) satisfy triangular inequality. Thus, the idea behind this approach is that given that the contact map is only an

approximation to the real distance matrix between protein residues (identifying only the shortest distances below a threshold), the best approximation to the full distance matrix (that satisfies the conditions for which the theorems of distance geometry hold) is guessed by the distance matrix computed from the shortest paths of the contact map.

The second step is based on the results of distance geometry (Sippl & Scheraga, 1985; Havel et al., 1983) and multidimensional scaling (Torgerson, 1952), which concern the reconstruction of the original spatial structure of a 3-dimensional object (in our case, the proteins) given the full distance matrix between its elements. This requires the spectral decomposition of the Gram matrix, defined according to the following formula:

$$G_{ij} = \frac{1}{2}\left[d_{0i}^2 + d_{0j}^2 - D_{ij}^2\right]$$

where

$$d_{0i}^2 = \frac{1}{N}\sum_{j=1}^{N} D_{ij}^2 - \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k>j}^{N} D_{jk}^2$$

represents the distance between the barycenter O and the point P_i of the 3D object. Cα spatial coordinates are then estimated through the 3 eigenvectors E_l ($l=1, 2, 3$) associated with the 3 largest eigenvalues λ_l ($l=1, 2, 3$) of the Gram matrix, as follows:

$$V_{li} = E_l(i) \times \sqrt{\lambda_l} \text{ with } \sum_{i=1}^{N} E_l^2(i) = 1.$$

where $E_l(i)$ is the i-th component of the eigenvector E_l. In this way, we can obtain a 3D reconstruction of Cα spatial coordinates up to an arbitrary rotation, dilation and possibly mirror symmetry (Lesne et al., 2014).

The two approaches start from different assumptions, but they rely on a similar algebraic structure, since also the Laplacian operator can be seen as a Gram matrix $L=I^TI$, where I is a rectangular incidence matrix that has one row for each link of the network, containing -1 and 1 values in each row in correspondence to the connected nodes (the direction of the link can be arbitrarily chosen, since this direction information is lost in the Laplacian operator).

We reconstructed the 3D structure of 63 proteins from (Menichetti et al., 2016) and characterized by different sizes (from 20 to 8015 Cα) and different

folding kinetics. The results were evaluated by computing the Pearson's correlation between Cα-pairwise distances in the reconstructed and real structure: the higher the correlation, the better the reconstruction.

We tested whether and how the following parameters could affect the results: (1) the folding kinetics; (2) the number of Cα composing the proteins. Moreover, the threshold value used to compute the contact map was varied, considering in particular 8 and 12 Å, to see how the variation in the resulting contact map could affect the reconstruction performance.

Looking at the histograms of correlation values represented in Fig. 1, we can notice that the reconstruction through ShRec3D achieved higher performances than the Laplacian-based one. In particular, the former is characterized by similar performances independently on the threshold value used to calculate the contact map, while the latter is characterized by an overall worsening for contact maps calculated using 12 Å as threshold (see Table 1). This result justifies the hypothesis that the shortest path distance matrix provides a more reliable estimation of the original distance matrix than the simple contact map, showing an increasing performance as the correlation value between shortest path distances and true distances increases (see Fig. 2). In fact, if we represent in a scatterplot the former as a function of the latter for each pair of Cα composing a protein, we can see that the structures that are well reconstructed by ShRec3D, show a linear

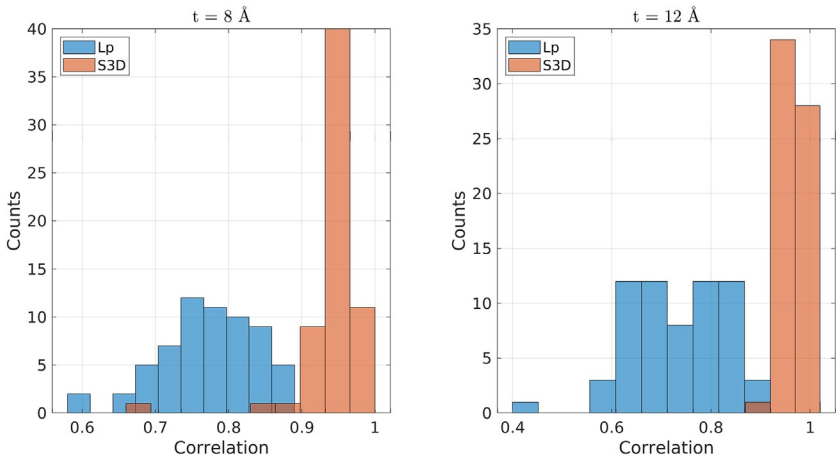

Fig. 1 Histograms of correlation values between real and reconstructed Cα-pairwise distances via Laplacian and ShRec3D embedding, starting from contact maps obtained using different threshold values: 8 and 12 Å.

Table 1 Mean correlation values between real and reconstructed Cα distances via Laplacian and ShRec3D method, starting from different contact maps, obtained using as threshold values 8 and 12 Å.

	$t = 8$ Å	$t = 12$ Å
Laplacian	0.77	0.73
ShRec3D	0.94	0.97

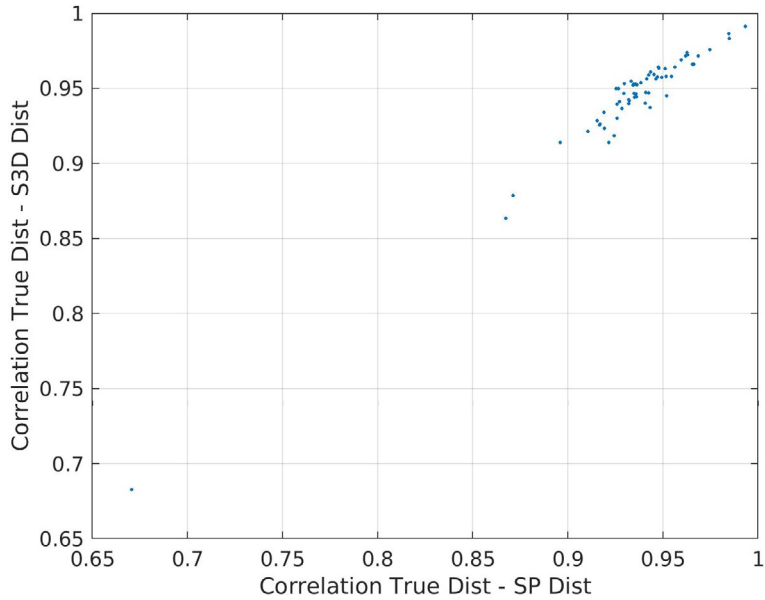

Fig. 2 Scatter plot between correlation values obtained by comparing real and reconstructed distances with Shrec3D (S3D Dist) and correlation values obtained by comparing real distances and shortest-path lengths (SP Dist), starting from an 8 Å contact map.

increasing trend characterized by a lower dispersion (see Fig. 3); whereas the structures that are not well reconstructed are characterized by a higher dispersion (see Fig. 4).

If we consider the 10 proteins with the highest and the lowest correlation values between real and reconstructed distances obtained from an 8 Å contact map, we can notice that Laplacian embedding provides the best results on proteins that do not show a modular structure (see Table 2, Figs. 5 and 6), which corresponds to a contact map characterized by blocks

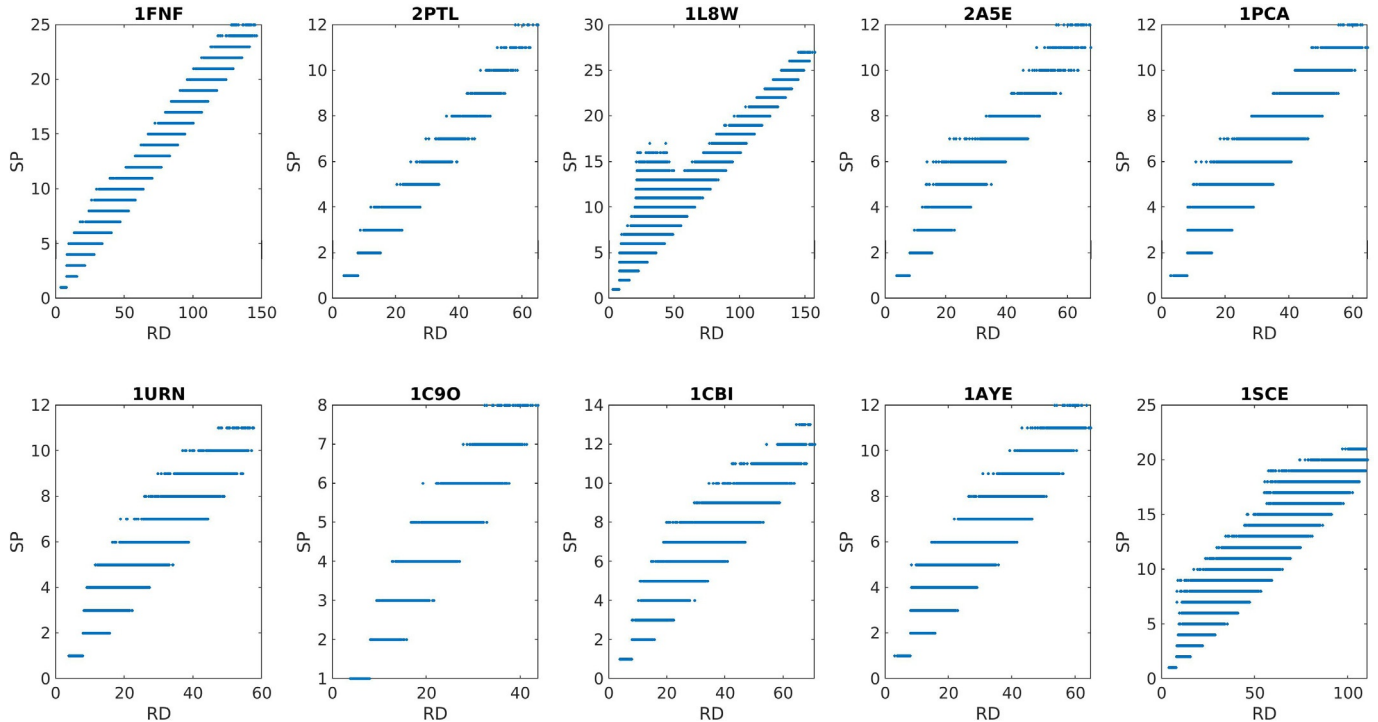

Fig. 3 Scatter plot between real distances (RD) and shortest-path lengths (SP) for the 10 proteins listed in Table 3, on which the *ShRec3D*-based method obtained the *best results*.

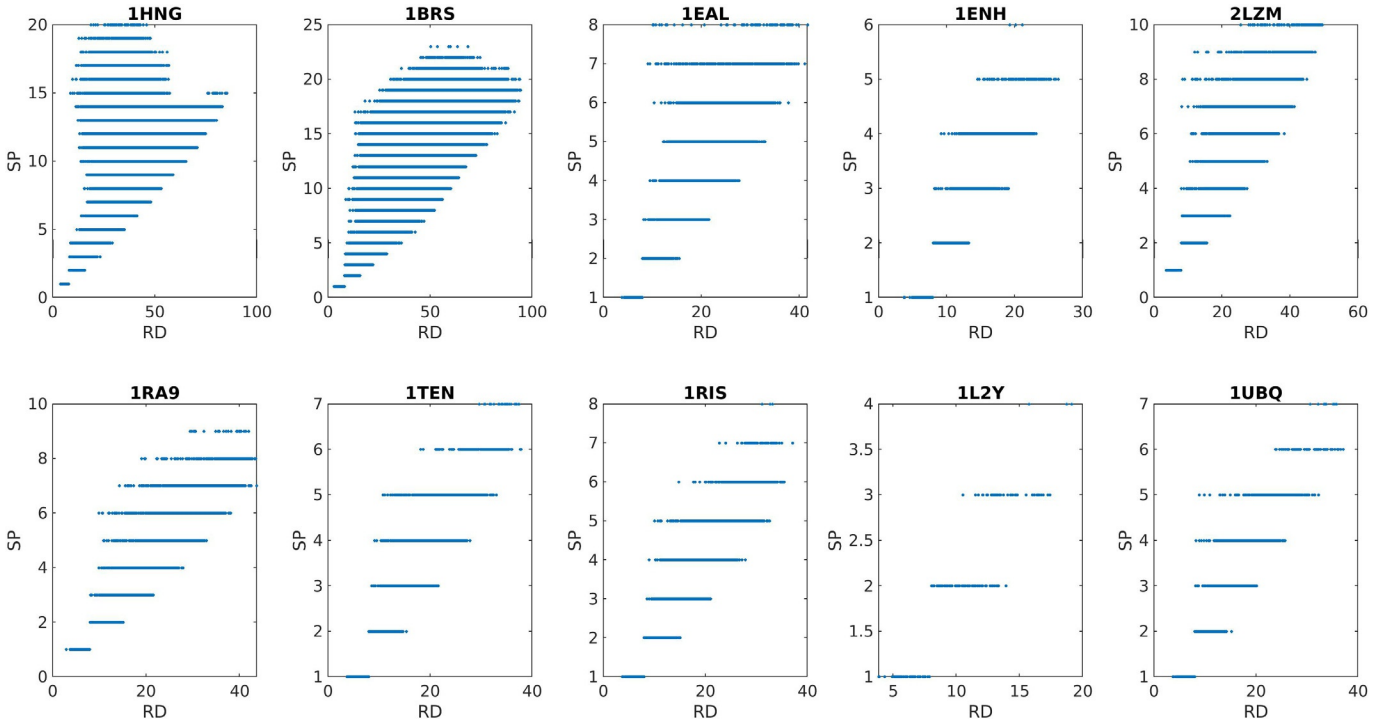

Fig. 4 Scatter plot between real distances (RD) and shortest-path lengths (SP) for the 10 proteins listed in Table 3, on which the *ShRec3D*-based method obtained the *worst results*.

Table 2 *Left*: 10 proteins with the *highest correlation values* between real distances and reconstructed ones via *Laplacian eigenvectors*, starting from an 8 Å contact map. All the proteins are characterized by the absence of a modular structure (MS, shown in Fig. 5 and even more clearly in Fig. 6) and 8 out of 10 belong to the two-state class (FK). *Right*: 10 proteins with the *lowest correlation values* between real distances and reconstructed ones via *Laplacian eigenvectors*, starting from an 8 Å contact map. 8 out of 10 proteins are characterized by a modular structure (MS, shown in Fig. 7 and even more clearly in Fig. 8) and 6 out of 10 belong to the multi-state class (FK).

Protein ID	MS	FK	Protein ID	MS	FK
2CRO	No	Multi-state	1SCE	Yes	Multi-state
1ARR	No	Two-state	1CBI	Yes	Multi-state
1BNZ	No	Two-state	1PBA	No	Two-state
1SRL	No	Two-state	1PHP	Yes	Multi-state
2PTL	No	Two-state	1FNF	Yes	Multi-state
1HRC	No	Two-state	1VII	No	Two-state
1BTA	No	Multi-state	1OPA	Yes	Multi-state
1POH	No	Two-state	1PIN	Yes	Two-state
1YCC	No	Two-state	2LZM	Yes	Multi-state
2ACY	No	Two-state	1C9O	Yes	Two-state

along the diagonal; on the contrary, ShRec3D embedding provides the worst results on proteins characterized by the absence of a modular structure (see Table 3, Figs. 11 and 12). In particular, the 10 proteins with the lowest correlation values between real and Laplacian-reconstructed distances and the 10 proteins with the highest correlation values between real and ShRec3D-reconstructed distances, have four elements in common: 1SCE, 1CBI, 1FNF, and 1C9O, which are all characterized by a contact map with blocks along the diagonal.

If we stratify the dataset according to protein two-state or multi-state folding kinetics, we can see that ShRec3D reaches the best performance when using 12 Å as threshold, independently on the two-state or multi-state class (see Table 4), whereas the Laplacian-based method reaches the best performance on two-state and multi-state proteins at different threshold values, which are respectively 8 and 12 Å (see Table 5).

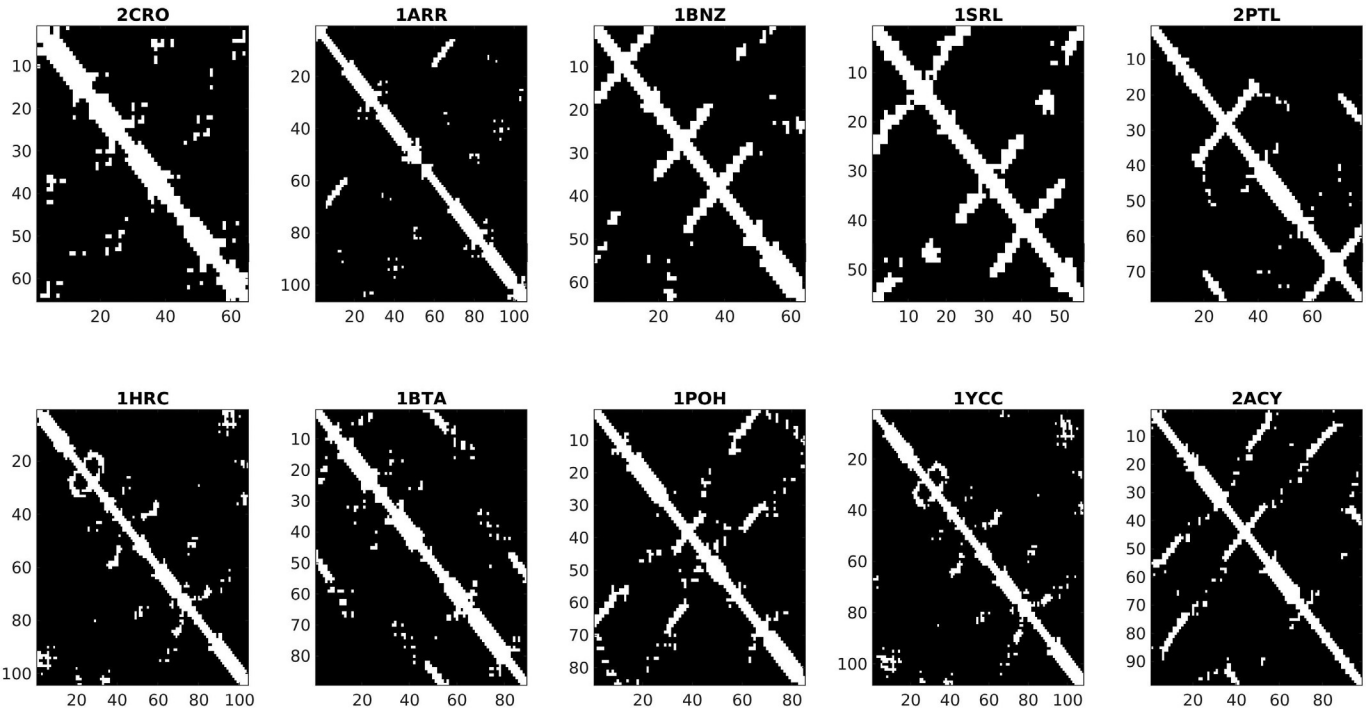

Fig. 5 8 Å contact maps of the 10 proteins listed in Table 2, with the *highest correlation values* between real distances and reconstructed ones via *Laplacian eigenvectors*.

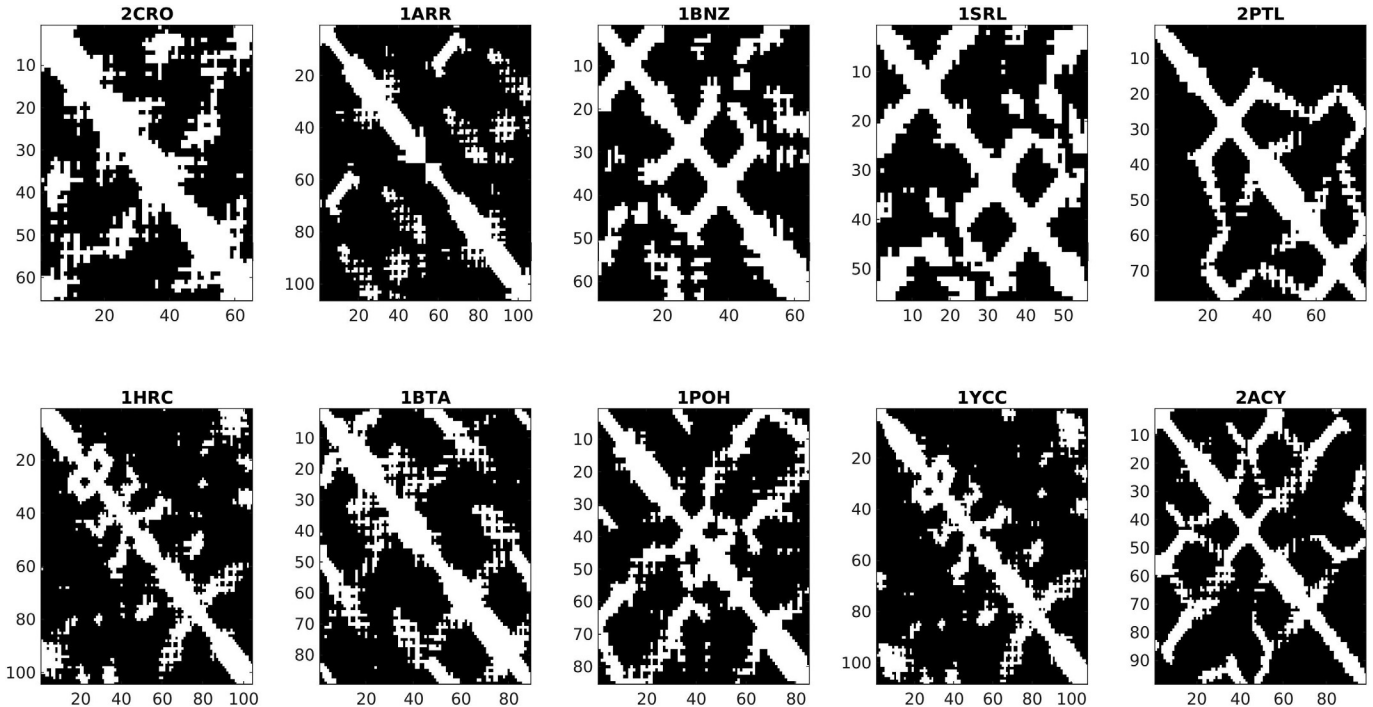

Fig. 6 12 Å contact maps of the 10 proteins listed in Table 2, with the *highest correlation values* between real and reconstructed distances via *Laplacian eigenvectors*.

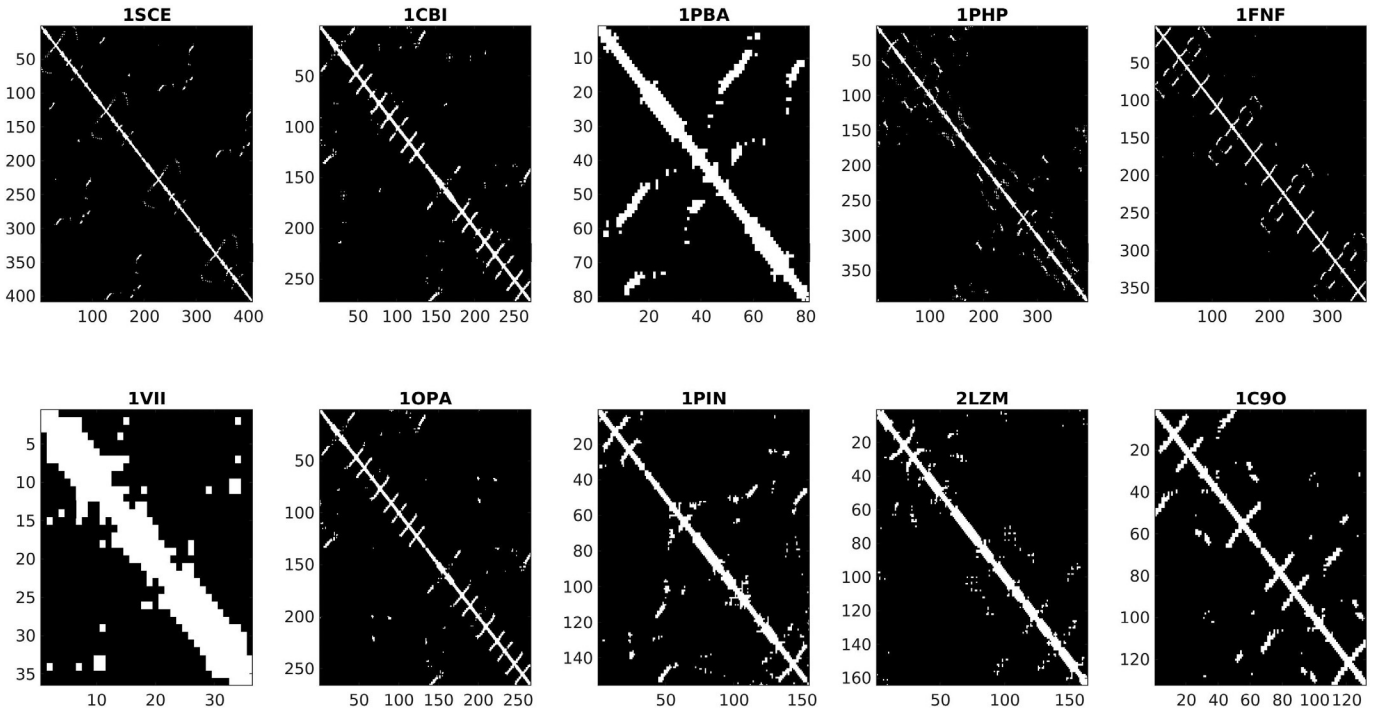

Fig. 7 8 Å contact maps of the 10 proteins listed in Table 2, with the *lowest correlation values* between real distances and reconstructed ones via *Laplacian eigenvectors*.

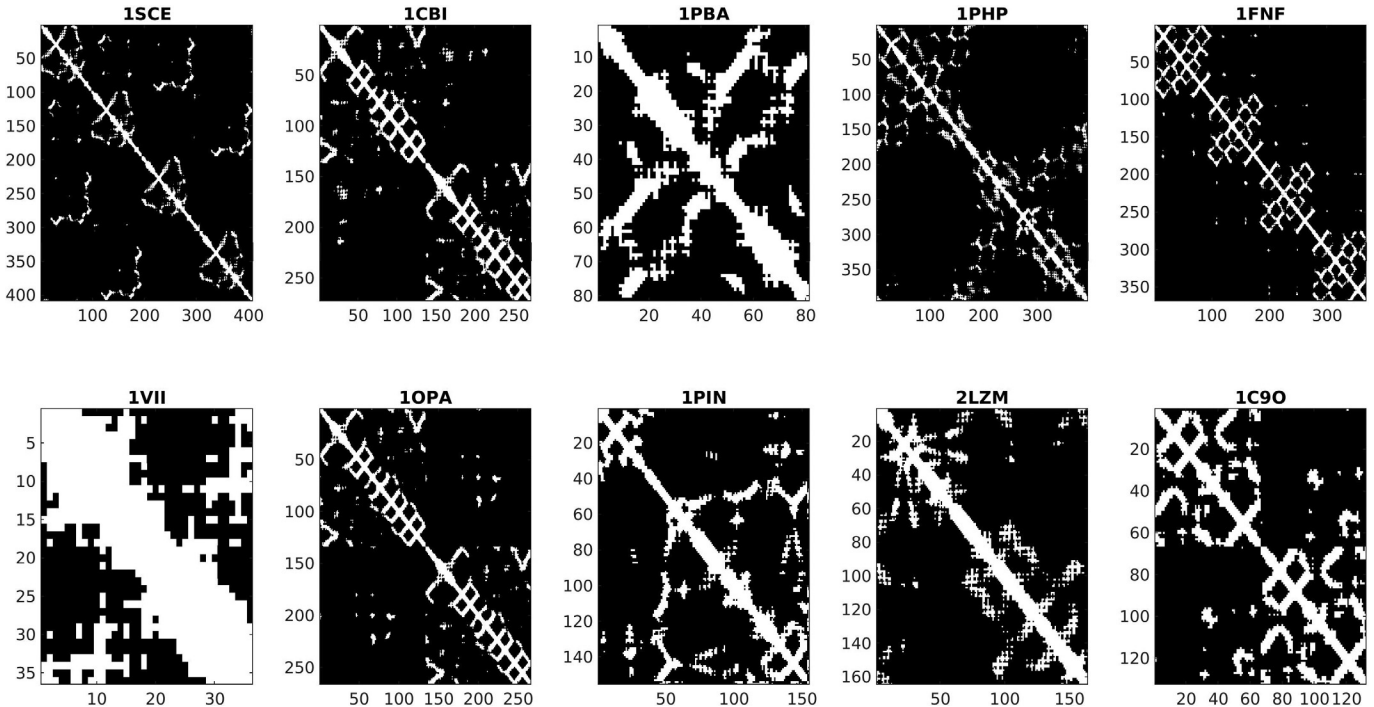

Fig. 8 12 Å contact maps of the 10 proteins listed in Table 2, with the *lowest correlation values* between real and reconstructed distances via *Laplacian eigenvectors*.

Table 3 *Left*: 10 proteins with the *highest correlation values* between real distances and reconstructed ones via *ShRec3D*, starting from an *8 Å contact map*. 6 out of 10 proteins are characterized by a modular structure (MS, shown in Fig. 9 and even more clearly in Fig. 10) and 7 out of 10 belong to the two-state class (FK). *Right*: 10 proteins with the *lowest correlation values* between real distances and reconstructed ones via *ShRec3D*, starting from an 8 Å contact map. 8 out of 10 proteins are characterized by the absence of a modular structure (MS, shown in Fig. 11 and even more clearly in Fig. 12) and 5 out of 10 belong to the multi-state class (FK).

Protein ID	MS	FK	Protein ID	MS	FK
1FNF	Yes	Multi-state	1HNG	Yes	Multi-state
2PTL	No	Two-state	1BRS	Yes	Multi-state
1L8W	Yes	Two-state	1EAL	No	Multi-state
2A5E	No	Multi-state	1ENH	No	Two-state
1PCA	No	Two-state	2LZM	No	Multi-state
1URN	Yes	Two-state	1RA9	No	Multi-state
1C9O	Yes	Two-state	1TEN	No	Two-state
1CBI	Yes	Multi-state	1RIS	No	Two-state
1AYE	No	Two-state	1L2Y	No	Two-state
1SCE	Yes	Multi-state	1UBQ	No	Two-state

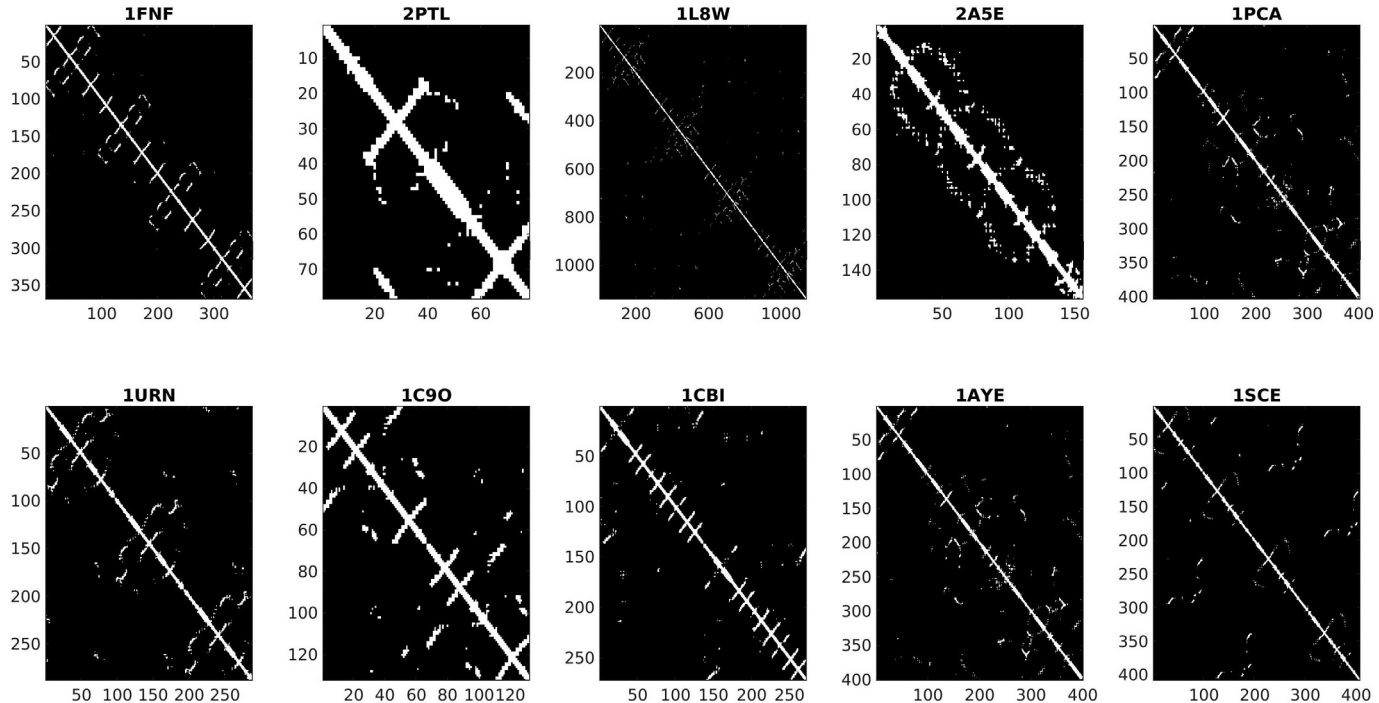

Fig. 9 8 Å contact maps of the 10 proteins listed in Table 3, with the *highest correlation values* between real distances and reconstructed ones via *ShRec3D*.

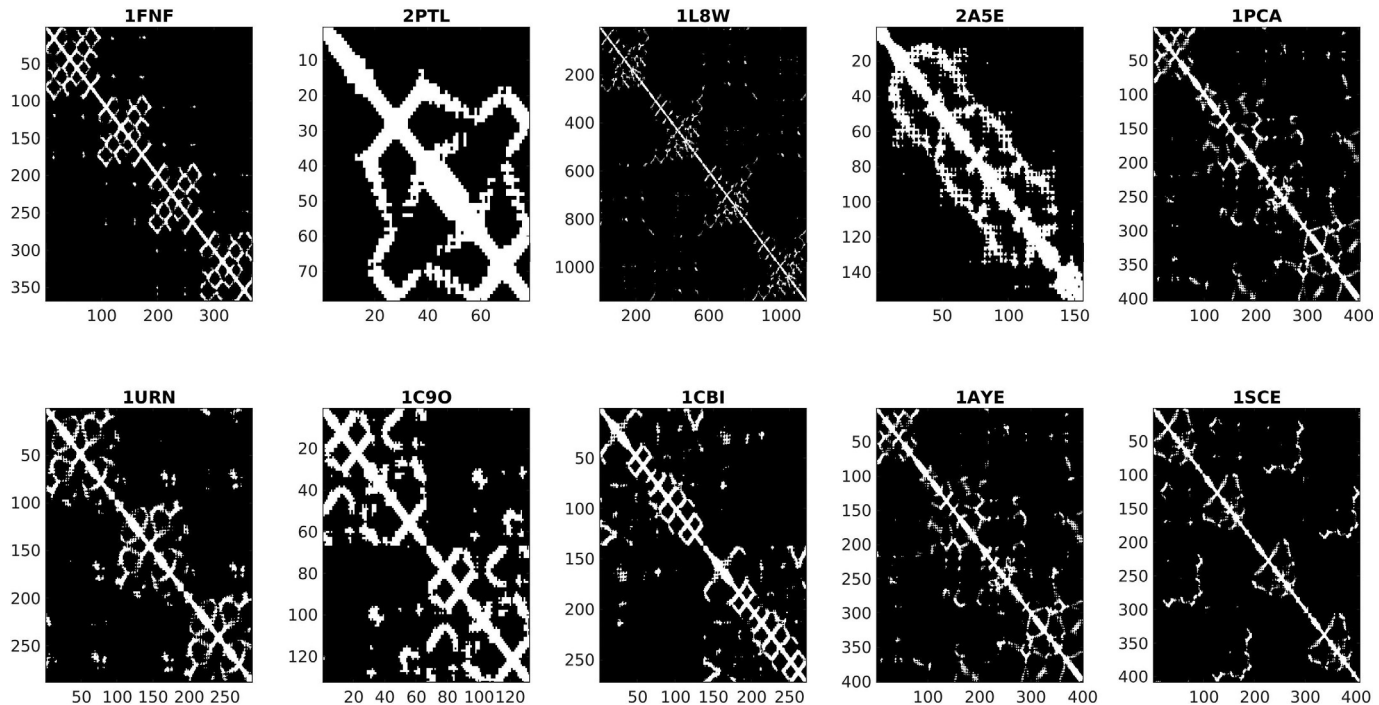

Fig. 10 12 Å contact maps of the 10 proteins listed in Table 3, with the *highest correlation values* between real and reconstructed distances via *ShRec3D*.

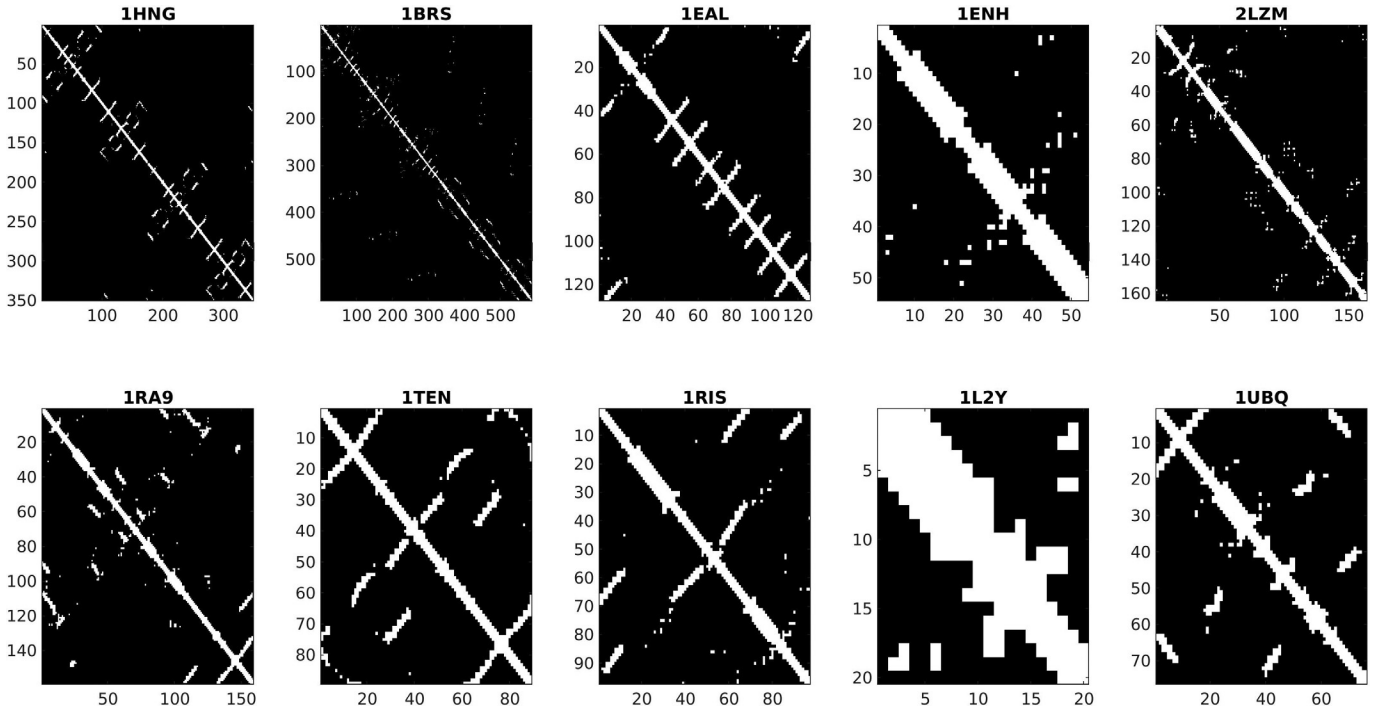

Fig. 11 8 Å contact maps of the 10 proteins listed in Table 3, with the *lowest correlation values* between real distances and reconstructed ones via *ShRec3D*.

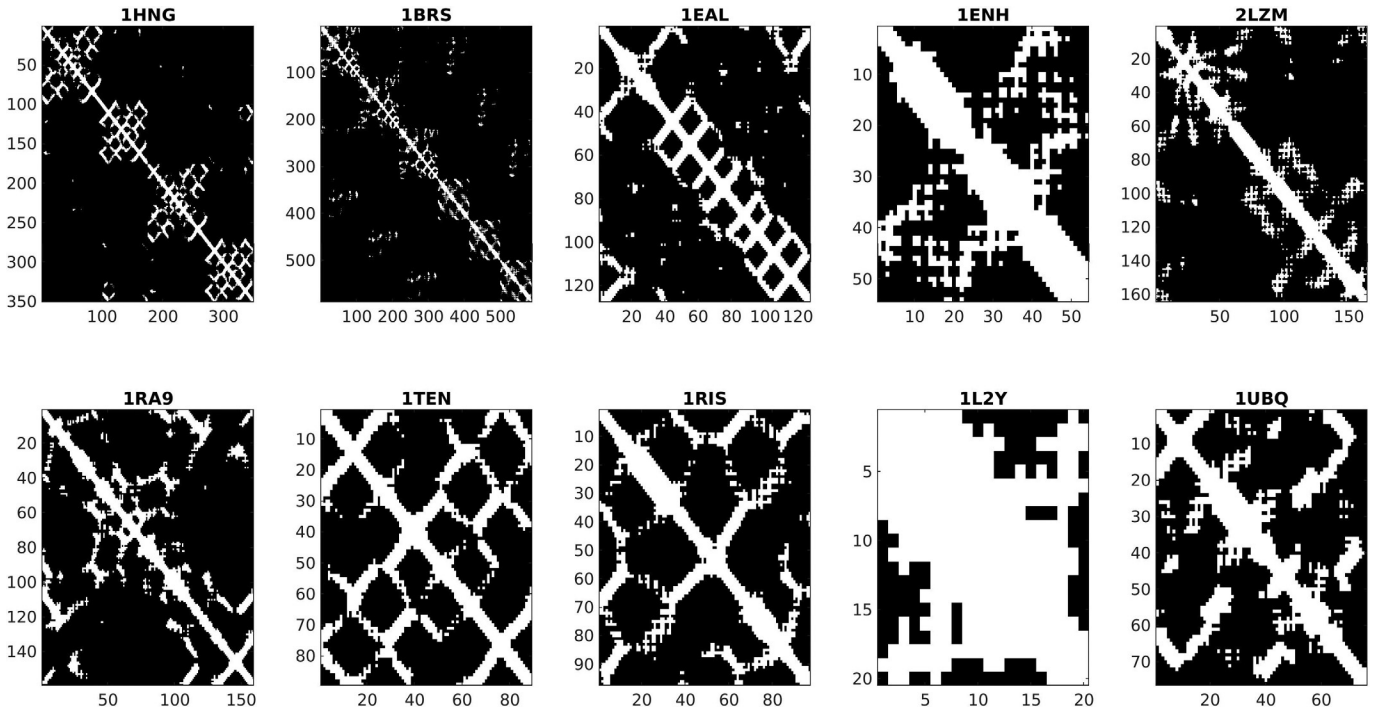

Fig. 12 12 Å contact maps of the 10 proteins listed in Table 3, with the *lowest correlation values* between real and reconstructed distances via *ShRec3D*.

Table 4 Mean correlation values between real and reconstructed Cα distances via *ShRec3D* method, for proteins divided into two-state or multi-state and represented by different contact maps, obtained using as threshold values 8 and 12 Å.

	Two-state	Multi-state
t = 8 Å	0.95	0.93
t = 12 Å	0.96	0.97

Table 5 Mean correlation values between real and reconstructed Cα distances via *Laplacian* method, for proteins divided into two-state or multi-state and represented by different contact maps, obtained using as threshold values 8 and 12 Å.

	Two-state	Multi-state
t = 8 Å	0.79	0.75
t = 12 Å	0.71	0.78

Moreover, if we represent the correlation values between real and reconstructed Cα distances as a function of the number of residues (see Figs. 13–16), we can notice two different behaviors between the two-state/multi-state classes, depending on the threshold used to calculate the contact map and on the reconstruction method. In fact, if we start from contact maps produced using 8 Å as threshold, we can see that the performance of both methods is almost constant as the length of the number of residues increases, for two-state proteins (see Fig. 15); whereas, in the case of multi-state proteins, this is still true only for ShRec3D reconstruction, while the Laplacian-based one shows a decreasing performance as the number of residues increases (see Fig. 15). If we start from contact maps produced using 12 Å as threshold, we can see that the scenario changes only for two-state proteins, whose performance increases as the number of residues increases, for both methods (see Fig. 16).

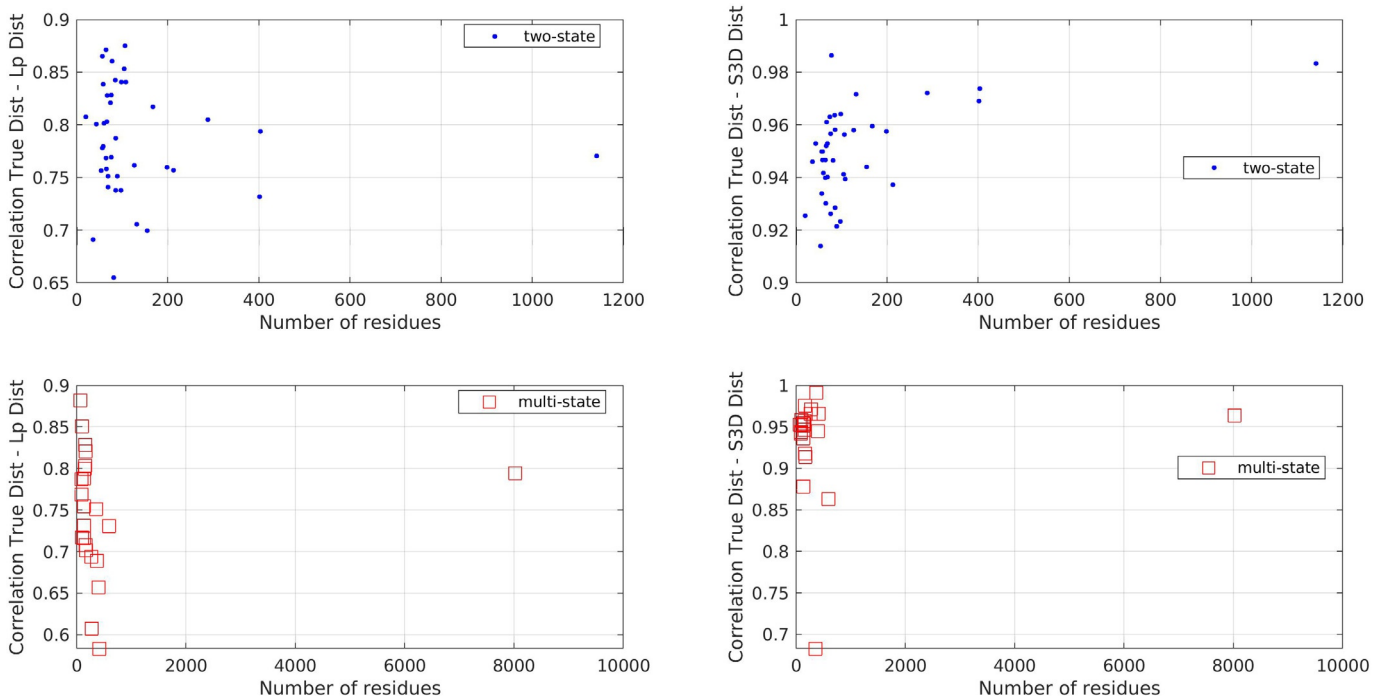

Fig. 13 Correlation value between true and reconstructed distances via Laplacian (left column) and ShRec3D (right column) embedding for two-state (upper row) and multi-state (lower row) proteins represented through an 8 Å contact map, as a function of the number of residues.

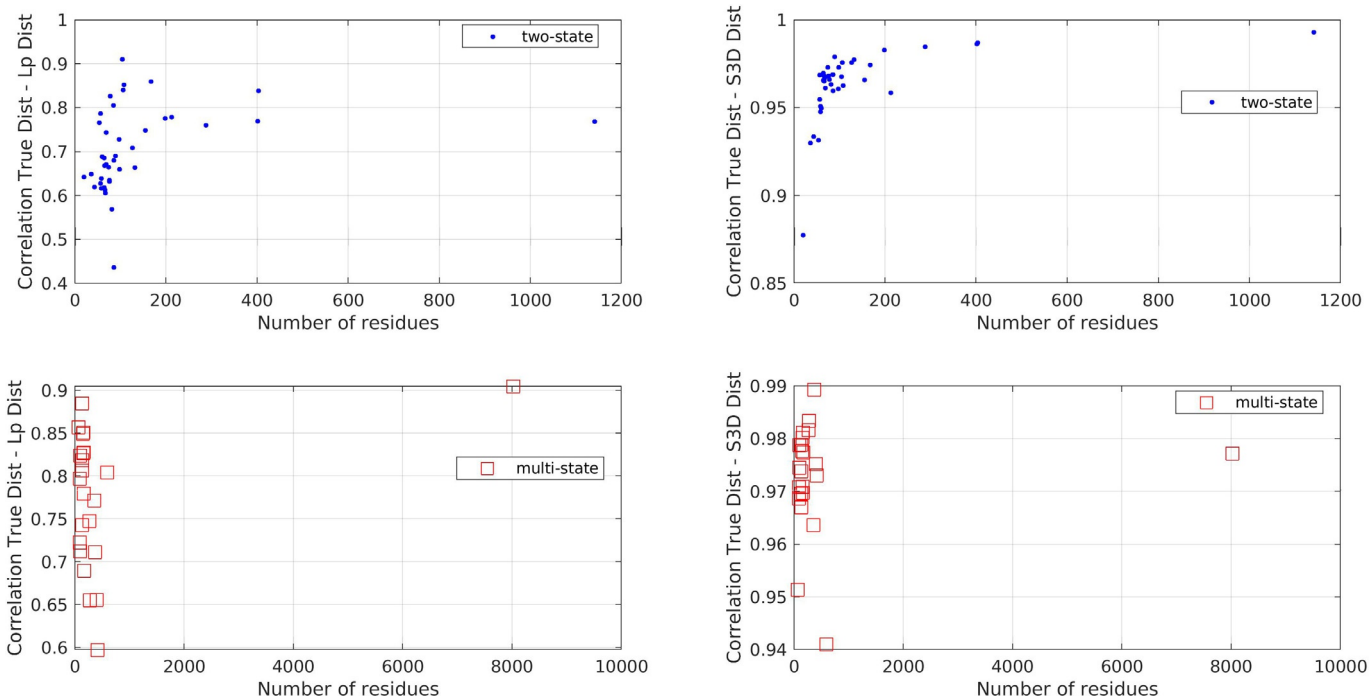

Fig. 14 Correlation value between true and reconstructed distances via Laplacian (left column) and ShRec3D (right column) embedding for two-state (upper row) and multi-state (lower row) proteins represented through a 12 Å contact map, as a function of the number of residues.

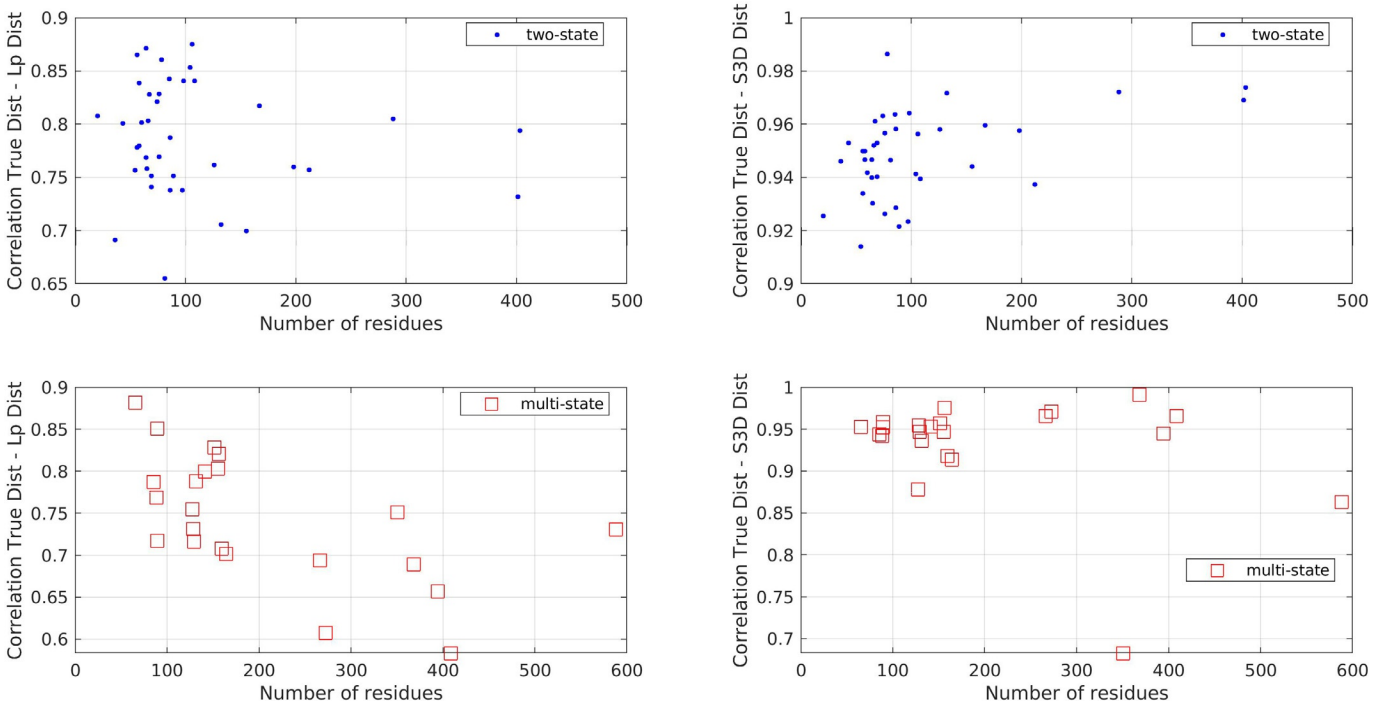

Fig. 15 Zoom of Fig. 10 on x-axis. Correlation value between true and reconstructed distances via Laplacian (left column) and ShRec3D (right column) embedding for two-state (upper row) and multi-state (lower row) proteins represented through an 8 Å contact map, as a function of the number of residues.

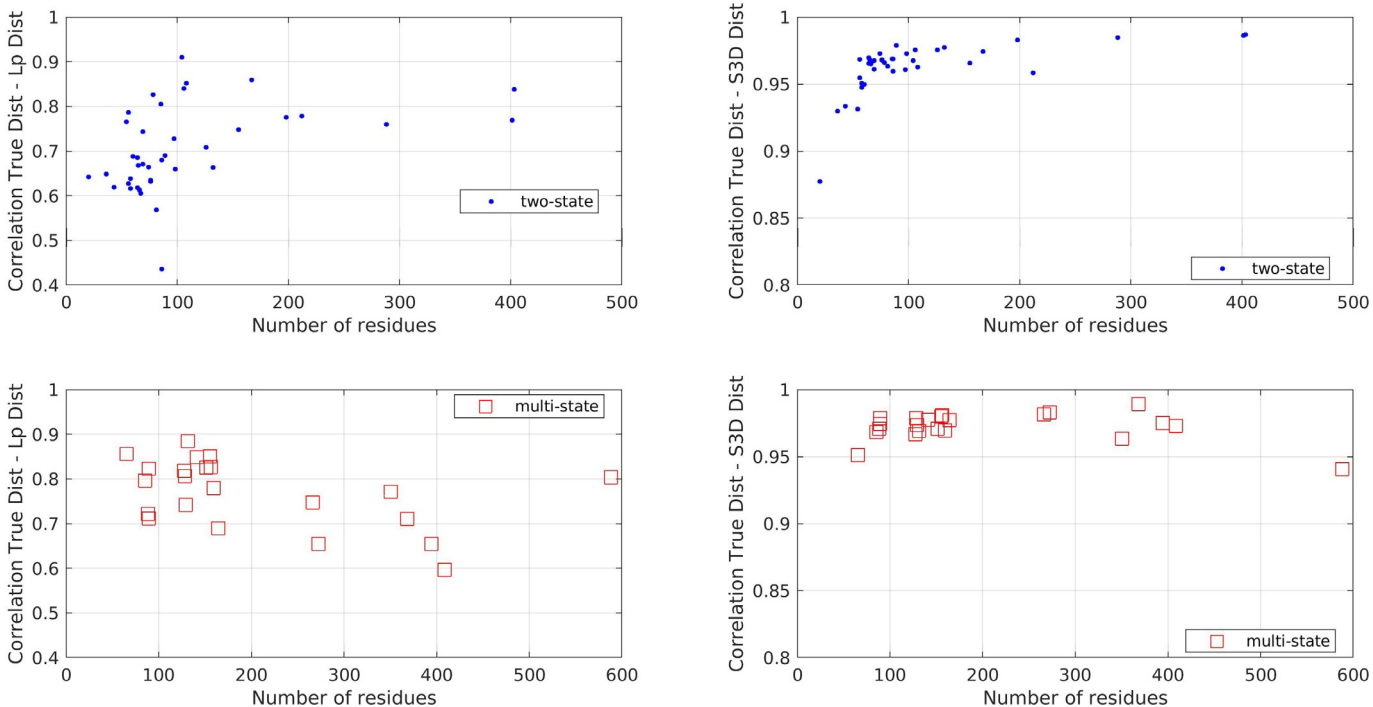

Fig. 16 Zoom of Fig. 11 on x-axis. Correlation value between true and reconstructed distances via Laplacian (left column) and ShRec3D (right column) embedding for two-state (upper row) and multi-state (lower row) proteins represented through a 12 Å contact map, as a function of the number of residues.

8. Discussion

Network-based methods can provide useful tools to characterize properties associated to protein folding, in particular regarding the 3D structure reconstruction. This scope can be interpreted as the identification of an optimal embedding manifold for the network through spectral approaches, related to the recently developing topic of network geometry (Boguñá et al., 2021), that deals with networks characterized by an intrinsic geometric space, in our case the 3D Euclidean space in which the 1D residue chain folds. At difference with less "physical" networks (like social networks, the world wide web, or protein interaction networks, in which there are no physical constraints on the links related to a maximum distance allowed between nodes) the properties of the chain structure of the protein, and the fact that it folds in a physical space, appear to be reflected in the properties of the contact map. In particular, the guess of using the shortest path distance as a proxy for the real residue distance, as proposed within the ShRec3D approach, seems satisfying in many cases, achieving better results than the Laplacian-based approach. As shown in a previous paper nonetheless (Menichetti et al., 2016), observables associated to the Laplacian operator allowed to discriminate between two-state and multi-state proteins, thus for other applications the informative content provided by this network formalism can be relevant as well. Even if the two proposed approaches rely on a common theoretical ground (i.e., the algebra of Gram matrices) the performances of the two methods are independent from each other (see Fig. 17) and seem to be much more influenced by the threshold value chosen for the computation of the contact map rather than the folding kinetics class (two-state/multi-state) or the number of residues. The database we used allowed us to evaluate and compare the two methods for the specific task of protein fold structure reconstruction, but in general the resulting spectral embeddings can be used on larger protein datasets as a pre-processing for unsupervised (i.e., clustering) or supervised (classification or mapping of specific protein chemical/physical properties) studies, and providing an optimal metrics for novel approaches like semi-supervised methods (van Engelen & Hoos, 2020).

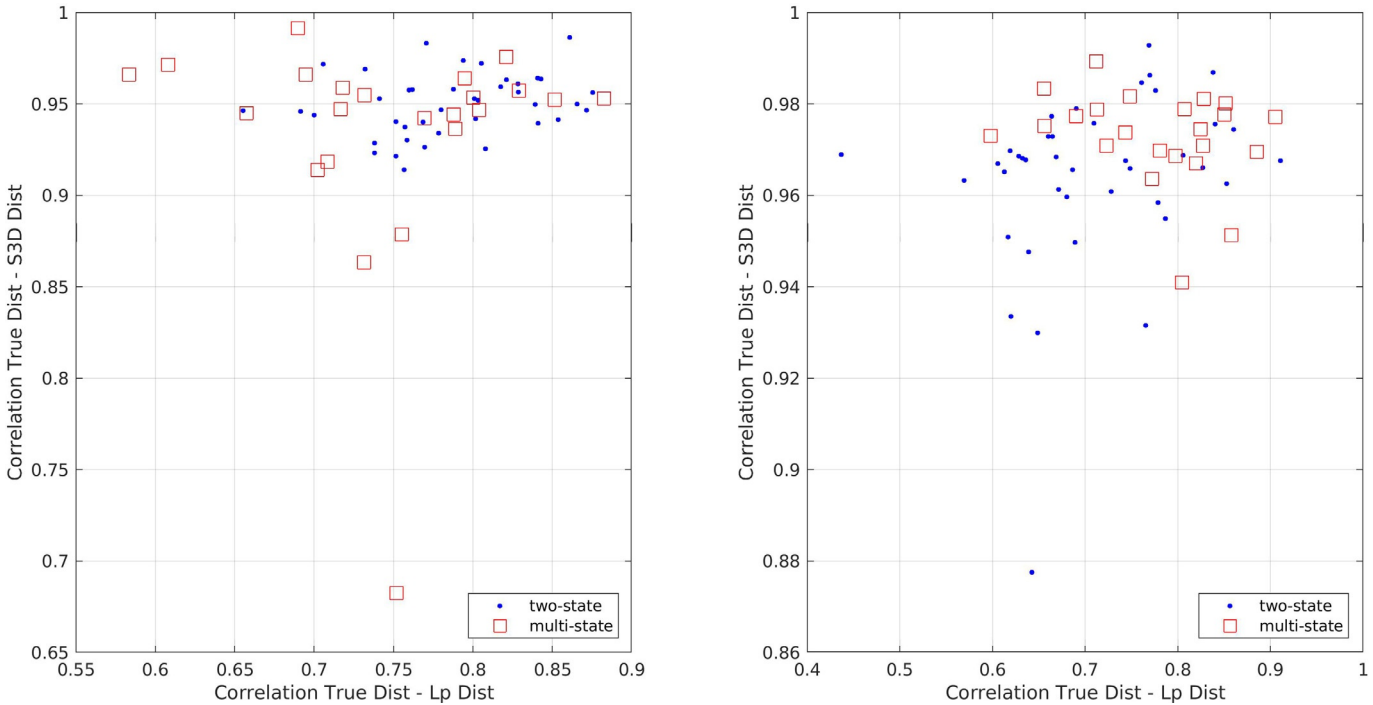

Fig. 17 Correlation value between true distances and reconstructed distances via ShR3c3D method as a function of correlation value between true distances and reconstructed distances via Laplacian-based method, obtained starting from an 8 Å contact map (left) and a 12 Å contact map (right).

References

Amitai, G., et al. (2004). Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, *344*, 1135–1146.

Andreeva, A., et al. (2020). The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, *48*, D376–D382.

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

Bagler, G., & Sinha, S. (2007). Assortative mixing in Protein Contact Networks and protein folding kinetics. *Bioinformatics*, *23*, 1760–1767.

Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, *294*, 93–96.

Bartoli, L., et al. (2008). The pros and cons of predicting protein contact maps. *Methods in Molecular Biology*, *413*, 199–217.

Biyikoglu, T., et al. (2007). *Laplacian eigenvectors of graphs: Perron-Frobenius and Faber-Krahn type theorems*. Berlin Heidelberg: Springer-Verlag.

Böde, C., et al. (2007). Network analysis of protein dynamics. *FEBS Letters*, *581*, 2776–2782.

Boguñá, M., et al. (2021). Network geometry. *Nature Reviews Physics*, *3*, 114–135.

Bollobas, B. (1998). *Modern graph theory*. New York: Springer-Verlag.

Brinda, K. V., & Vishveshwara, S. (2005). A network representation of protein structures: Implications for protein stability. *Biophysical Journal*, *89*, 4159–4170.

Capriotti, E., & Casadio, R. (2007). K-Fold: A tool for the prediction of the protein folding kinetic order and rate. *Bioinformatics*, *23*, 385–386.

Chang, C. C. H., et al. (2015). Towards more accurate prediction of protein folding rates: A review of the existing Web-based bioinformatics approaches. *Briefings in Bioinformatics*, *16*, 314–324.

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*, 823–826.

Compiani, M., & Capriotti, E. (2013). Computational and theoretical methods for protein folding. *Biochemistry*, *52*, 8601–8624.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, *29*, 7133–7155.

Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. *Science*, *338*, 1042–1046.

Dill, K. A., et al. (2007). The protein folding problem: When will it be solved? *Current Opinion in Structural Biology*, *17*, 342–346.

Dishman, A. F., & Volkman, B. F. (2018). Unfolding the mysteries of protein metamorphosis. *ACS Chemical Biology*, *13*, 1438–1446.

Fariselli, P., et al. (2007). The WWWH of remote homolog detection: The state of the art. *Briefings in Bioinformatics*, *8*, 78–87.

Grabowski, M., et al. (2016). The impact of structural genomics: The first quindecennial. *Journal of Structural and Functional Genomics*, *17*, 1–16.

Greene, L. H. (2012). Protein structure networks. *Briefings in Functional Genomics*, *11*, 469–478.

Grewal, R. K., & Roy, S. (2015). Modeling proteins as residue interaction networks. *Protein and Peptide Letters*, *22*, 923–933.

Gromiha, M. M., & Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. *Journal of Molecular Biology*, *310*, 27–32.

Havel, T. F., et al. (1983). The theory and practice of distance geometry. *Bulletin of Mathematical Biology*, *45*, 665–720.

Hrmova, M., & Fincher, G. B. (2009). Functional genomics and structural biology in the definition of gene function. *Methods in Molecular Biology*, *513*, 199–227.

Huang, L.-T., & Gromiha, M. M. (2010). First insight into the prediction of protein folding rate change upon point mutation. *Bioinformatics*, *26*, 2121–2127.

Ivankov, D. N., & Finkelstein, A. V. (2004). Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 8942–8944.
Karplus, M., & Weaver, D. L. (1976). Protein-folding dynamics. *Nature*, *260*, 404–406.
Kryshtafovych, A., et al. (2019). Critical assessment of methods of protein structure prediction (CASP)-Round XIII. *Proteins*, *87*, 1011–1020.
Lesne, A., et al. (2014). 3D genome reconstruction from chromosomal contacts. *Nature Methods*, *11*, 1141–1143.
Lieberman-Aiden, E., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*, 289–293.
Magliery, T. J. (2015). Protein stability: Computation, sequence statistics, and new experimental methods. *Current Opinion in Structural Biology*, *33*, 161–168.
Menichetti, G., et al. (2016). Network measures for protein folding state discrimination. *Scientific Reports*, *6*, 30367.
Merlotti, A., et al. (2020). Merging 1D and 3D genomic information: Challenges in modelling and validation. *Biochimica et Biophysica Acta, Gene Regulatory Mechanisms*, *1863*, 194415.
Plaxco, K. W., et al. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *Journal of Molecular Biology*, *277*, 985–994.
Porto, M., et al. (2004). Reconstruction of protein structures from a vectorial representation. *Physical Review Letters*, *92*, 218101.
Punta, M., & Rost, B. (2005). Protein folding rates estimated from contact predictions. *Journal of Molecular Biology*, *348*, 507–512.
Sanavia, T., et al. (2020). Limitations and challenges in protein stability prediction upon genome variations: Towards future applications in precision medicine. *Computational and Structural Biotechnology Journal*, *18*, 1968–1979.
Senior, A. W., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, *577*, 706–710.
Sillitoe, I., et al. (2021). CATH: Increased structural coverage of functional space. *Nucleic Acids Research*, *49*, D266–D273.
Sippl, M. J., & Scheraga, H. A. (1985). Solution of the embedding problem and decomposition of symmetric matrices. *Proceedings of the National Academy of Sciences of the United States of America*, *82*, 2197–2201.
Taylor, N. R. (2013). Small world network strategies for studying protein structures and binding. *Computational and Structural Biotechnology Journal*, *5*, e201302006.
Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*, 401–419.
van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*, 373–440.
Vassura, M., et al. (2008). Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *5*, 357–367.
wwPDB consortium. (2019). Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, *47*, D520–D528.
Yan, W., et al. (2014). The construction of an amino acid network for understanding protein structure and function. *Amino Acids*, *46*, 1419–1439.
Zhou, H., & Zhou, Y. (2002). Folding rate prediction using total contact distance. *Biophysical Journal*, *82*, 458–463.