Coevolution and smFRET Enhances Conformation Sampling and FRET Experimental Design in Tandem PDZ1−2 Proteins

Published as part of The Journal of Physical Chemistry virtual special issue "Jose Onuchic Festschrift".

Aishwarya Krishnamohan,[||] George L. Hamilton,[||] Rajen Goutam, Hugo Sanabria,* and Faruck Morcos*
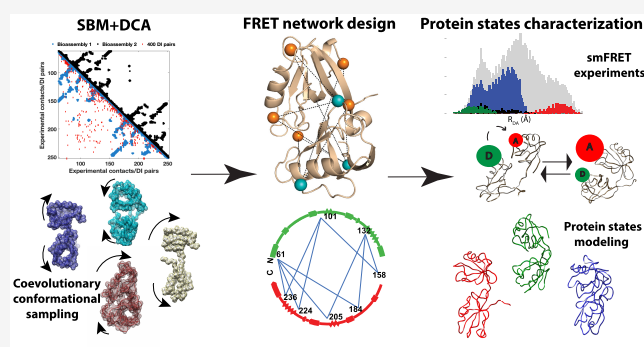
ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: The structural flexibility of proteins is crucial for their functions. Many experimental and computational approaches can probe protein dynamics across a range of time and length-scales. Integrative approaches synthesize the complementary outputs of these techniques and provide a comprehensive view of the dynamic conformational space of proteins, including the functionally relevant limiting conformational states and transition pathways between them. Here, we introduce an integrative paradigm to model the conformational states of multidomain proteins. As a model system, we use the first two tandem PDZ domains of postsynaptic density protein 95. First, we utilize available sequence information collected from genomic databases to identify potential amino acid interactions in the PDZ1−2

tandem that underlie modeling of the functionally relevant conformations maintained through evolution. This was accomplished through combination of coarse-grained structural modeling with outputs from direct coupling analysis measuring amino acid coevolution, a hybrid approach called SBM+DCA. We recapitulated five distinct, experimentally derived PDZ1−2 tandem conformations. In addition, SBM+DCA unveiled an unidentified, twisted conformation of the PDZ1−2 tandem. Finally, we implemented an integrative framework for the design of single-molecule Förster resonance energy transfer (smFRET) experiments incorporating the outputs of SBM+DCA with simulated FRET observables. This resulting FRET network is designed to mutually resolve the predicted limiting state conformations through global analysis. Using simulated FRET observables, we demonstrate that structural modeling with the newly designed FRET network is expected to outperform a previously used empirical FRET network at resolving all states simultaneously. Integrative approaches to experimental design have the potential to provide a new level of detail in characterizing the evolutionarily conserved conformational landscapes of proteins, and thus new insights into functional relevance of protein dynamics in biological function.

INTRODUCTION

Proteins are flexible biomolecules that can adopt multiple structural conformations over varied length scales and time scales.[1] The structural dynamics of proteins are often crucial to their biological functions. Intrinsically disordered proteins and protein regions represent the extreme of this phenomenon, often containing no discernible secondary or tertiary fold but performing varied functions.[2−4] Larger proteins often contain several folded domains connected by flexible disordered linkers, giving rise to distinct conformations in which the domains reorient relative to one another[5,6] and thus to the supertertiary structure.[7] Such is the case for the post synaptic density protein of 95 kDa (PSD-95),[6] a key protein that underlies the organization of postsynaptic proteins by bringing together multiple binding partners. PSD-95 contains five domains that dynamically exchange between multiple configurations.[5,6,8] The dynamic behaviors of proteins are complicated further by the adoption of higher-order protein

complexes.[9] Several computational and experimental techniques are available to investigate the structure and dynamics of protein systems. Molecular dynamics (MD) simulations can provide an atomistic view into the structural dynamics of biomolecules. However, computational costs make capturing large or slow conformational changes challenging or impractical.[10,11] Experimentally, X-ray crystallography[12] and nuclear magnetic resonance (NMR)[13] can resolve very high-resolution, atomistic protein structures. However, ensemble and time averaging of experimental observables results in loss
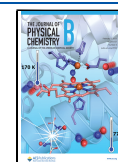
of information about dynamics, or even structures that do not represent the functional limiting states of a structurally dynamic protein but instead averages of multiple distinct conformations.

Single-molecule techniques aim to avoid such ensemble averaging by probing the characteristics of individual molecules. Single-molecule Förster resonance energy transfer (smFRET) can be used to resolve distances between fluorophores tethered to biomolecular surfaces with Ångstrom-scale precision.[14] Further, smFRET can be used to characterize both the limiting conformational states of proteins and dynamic exchange between them when combined with complementary modeling techniques.[15−18] Integrative structural biology techniques seek to synthesize information from multiple computational and experimental techniques to obtain models of biomolecular systems that are consistent across the time and length scales probed by the individual techniques.[19−22] Integration of smFRET with other techniques in novel integrative approaches has revealed a wealth of information about dynamic biomolecular systems.[16,23−26] Recently, we used structural restraints from FRET experiments in conjunction with MD simulations to identify two dynamically exchanging conformations of the PDZ1−2 tandem of PSD-95 that can explain discrepancies between previously obtained structures as resulting from dynamic averaging.[8] Integrative approaches provide an attractive path forward for addressing these and other biological questions. Recently, a database for deposition of biomolecular structures resulting from hybrid and integrative studies was established as the wwPDB.[27] However, development and implementation of integrative methodologies present unique challenges for the accessibility and design of studies and interpretation of results. Integrative smFRET approaches have been made more accessible than ever thanks to open science practices in development of new techniques and open-source platforms for these studies.[15,16,28−30] To further address the challenges associated with designing and interpreting hybrid smFRET studies, here we explore the integration of protein sequence information as a means of predicting structural information for the design of smFRET experiments that can probe predicted structural conformations of proteins. We additionally present this approach as a general workflow and use simulated FRET data to obtain clearly defined and testable hypothetical results for the designed experiments.

The amino acid sequences of proteins encode for their structure, structural dynamics, and functions. Evolutionarily conserved and coevolving residue positions in protein sequences belonging to a particular family can thus provide key insights into the regions of that protein that are relevant to maintaining its structure and function during evolution. Direct coupling analysis (DCA) is a technique for quantifying the degree of direct relationship between pairs of residue positions in a protein sequence during evolution without influence from other amino acid positions.[31] For each pair of sites, a theoretical information score for amino acid coevolution, e.g., direct information (DI), for each pair of residues can then be used as predictors for direct contacts and interactions between amino acids that result both from stable protein native folds,[32−36] multiple alternative protein conformations,[11,37,38] and protein−protein interactions.[39−42] Here, we combine DCA-based restraints with coarse-grained physical models of proteins like structure-based modeling (SBM) simulations to identify novel potential conformations of the PDZ1−2 tandem

of PSD-95. We further integrate the outcomes of these simulations into a workflow for design of FRET experiments through selection of fluorophore labeling sites catered toward mutually resolving the known and predicted conformations of the PDZ1−2 tandem. While strides have been made recently,[29] the optimal selection of residues for placement of fluorophores remains an open challenge for design of FRET experiments. The prediction of novel states, combined with DI from DCA analysis, allows informed selection of target states as well as information about amino acids which are important for native conformations and function, and that should therefore be avoided in labeling site selection. Thus, the resulting FRET network, or set of FRET pair labeling sites, should be better suited both for simultaneously recapitulating the conformations of interest and for avoiding perturbations to the protein's native behavior. Therefore, the synthesis of SBM+DCA with smFRET may provide a powerful toolset for the identification and modeling of novel structural conformations and dynamics in diverse multidomain protein systems.

■ METHODS

Framework. We designed a framework illustrated (see the "Results and Discussion" section) applied it to the PDZ1−2 tandem. We provide a brief description of the framework and the details of each of these steps are explained successively. The MSA of the sequences of protein domain is extracted from the Pfam database,[43] DCA is applied to the MSA to calculate the parameters e_{ij} and h_i. These parameters are then used to score and determine the strong coevolutionary couplings, which are incorporated into the SBM energy function. The conformational transitions of a protein from these SBM+DCA simulations are used to predict the residue positions that could be involved in the dynamics of a protein. This knowledge is then applied to FRET network design and experiments.

Data Extraction and Processing. In this study, we extracted the sequence of the protein from Protein Data Bank (PDB) (PDB ID: 3GSL) and used it as a seed sequence (input) for an in-house bash script. Our script uses the input seed sequence (containing both PDZ1 and 2 domains) and generates the MSA for both PDZ1−2 based on HMMER[44,45] with a 20% gap filtering cut off, i.e., we filtered out the sequences with more than 20% of the sequence length L of the protein with contiguous gaps.[43] Since we require at least one structure to run the SBM+DCA model, the crystal structures of PDZ1−2 tandem of PSD-95 protein with resolution less than 3 Å were obtained from the PDB.[46]

Direct Coupling Analysis (DCA). PDZ domain family MSA is arranged in M rows and L columns where M is the number of sequences and L is the length of each sequence which is denoted by

$$A = A_i^x \quad i = (1, ..., L) \quad x = (1, ..., M) \tag{1}$$

DCA was applied to the MSA of PDZ domain family using this mean field version of DCA[31] using DCA web-based server or an in-house MATLAB script to estimate the directly coupled coevolving residue pairs based on the DI score. DI scores are estimated based on the effective number of sequences M_{eff} which is computed after down-weighting the sequences with 70% identity. These DI pairs from PDZ family MSA were aligned and mapped back to the residues of the PDZ1−2 tandem protein using an in-house mapping script to visualize the coevolutionary pairs in the protein structure.

SBM+DCA Model and Molecular Dynamics (MD) Simulations. In the SBM+DCA hybrid model, we combined the SBM energy function from a single experimental structure of the PDZ1−2 tandem (PDB ID: 3GSL) and incorporated the DCA couplings into this function.[47,48] The equation for this hybrid method is given by

$$V_{total} = V_b^{(SBM)} + V_{nb}^{(SBM+DCA)} \quad (2)$$

Here V_{total} is the summation of energies of local interactions in the single SBM derived from experimental structure $V_b^{(SBM)}$, and the second term $V_{nb}^{(SBM+DCA)}$ in eq 2 includes nonlocal interactions in the SBM and the DCA and the repulsion between the nonbonded pairs that are not in contact in SBM or DCA. The expansion for each term in eq 2 is given below:

$$V_b^{SBM} = \sum_{bonds} K_r(r - r_0)^2 + \sum_{angles} K_\theta(\theta - \theta_0)^2 +$$
$$\sum_{dihedrals} K_\varphi^{(n)}[1 + \cos(n(\varphi - \varphi_0))] \quad (3)$$

$$V_{nb}^{SBM+DCA} = \sum_{ij} V_{ij}^G(r_{ij})_{SBM} + \sum_{ij} V_{ij}^G(r_{ij})_{DCA}$$
$$+ \sum_{ij} V_{ij}^R(r_{ij})_{SBM+DCA} \quad (4)$$

In eq 3 the first and second terms ensure that the bond distance between two residues (r), angles between three subsequent residues (θ) of the system at every simulation step are constrained harmonically with the native structure's bond distance, angle value (r_0, θ_0) by a spring constant $K_r = 20$ (kJ/mol. Å^2), $K_\theta = 20$ (kJ/mol·rad^2). The third term gives the difference of the dihedral angle between four subsequent residues along the backbone at the native structure (φ_0) and the system at every simulation step, where the dihedral angle potential $K_\varphi^{(1)} = 2K_\varphi^{(3)}$. In eq 4, the first and second terms are Gaussian attractive potentials for nonlocal interactions between residues i and j in SBM and DCA, where the nonlocal contacts of DCA are the top ranked strong coevolutionary couplings. The third term in eq 4 denotes the repulsion potential of the nonbonded interactions present either in SBM or DCA. Here, we use the distance for DCA pairs $(r_{ij})_{DCA} = 7.5$ Å. The SBM topology files with potentials for MD simulations were built using the SMOG server[49] based on eqs 3 and 4, and we used a cut off threshold of 6 Å of maximum contacting distance +1 Å atom shadowing radius (standard value used in SMOG server). The potential files for DCA contacts (eq 4) were generated by the in-house python package called *sbm-tools* and integrated into the SBM potential files.

In this study, we generated C-α models with a Gaussian contact interaction.[50] For all the systems we studied using this model, we used the number of DCA contacts ranging from 1 to 2 times the number of native contacts. The SBM+DCA potential input files were used for MD simulations with the GROMACS software package.[51] The simulations were performed at the temperature $T = 70$ K (we chose this temperature to avoid large thermal fluctuations in the simulation) with 100 million integration steps to sample the equilibrium conformational ensemble. The simulation trajectory is visualized and analyzed using the VMD[52] and Chimera[53] software packages.

FRET Network Design Protocol. FRET network design proceeded through the following steps (Figure S1):

(1) Target conformational state structures are chosen. The close-like (CL), open-like (OL), Twisted, and Extended states were chosen as the target states for design of the FRET network for the PDZ1−2 tandem. The final FRET network should mutually resolve the target states.

(2) A candidate list of individual fluorophore labeling sites is generated. Initially, the list includes all amino acids in the protein. Then, the top DI pairs from DCA are eliminated along with all residues that participate in contacts with others in the target structures. In this work, the top 1000 DI pairs were removed, and the cutoff for inter-residue contacts was set to 5 Å. This is to ensure that fluorophore labeling does not perturb native contacts associated with any predicted conformation. Additional labeling sites can be eliminated based on *a priori* information, such as known binding sites not captured through DCA. For the PDZ1−2 tandem, residues 151−155, constituting the interdomain linker, were eliminated as the primary motions of interest are interdomain motions.

(3) Accessible volume (AV) simulations are performed for all remaining candidate labeling sites to obtain simulated FRET observables. Details for these simulations are provided below.

(4) Predicted FRET efficiencies for all pairs of candidate labeling sites were calculated for all target structures. FRET efficiency calculations utilized 10000 pairs of points sampled from the AVs of the fluorophores, from which distances were calculated and the averages were converted into FRET efficiency values using

$$E = \frac{1}{1 + \left(\frac{\langle R_{DA} \rangle}{R_0}\right)^6} \quad (5)$$

where $\langle R_{DA} \rangle$ is the mean interdye distance, with a 52 Å Förster radius (R_0), corresponding to the previously used Alexa-488/Alexa-647 FRET pair.[8] The vectors spanning the mean positions of the AVs for each candidate FRET pair were also calculated.

(5) For each FRET pair, the observable contrast between the FRET efficiency for each target state and each other target state was calculated, using the maximum dynamic shift[54] as the contrast metric. The maximum dynamic shift is calculated via

$$S = \frac{1}{\sqrt{2}}(\sqrt{1 - E_1} - \sqrt{1 - E_2})^2 \quad (6)$$

where E_1 and E_2 are the FRET efficiency values of the compared conformations for a given FRET pair. The maximum dynamic shift describes the deviation of FRET observables from the so-called static FRET line when transitions between two states with different FRET efficiencies occur during the observation times of the molecules.

(6) The FRET pair with the highest average maximum dynamic shift among all target states (P101−Y236 for the target states used here) was used as the first FRET pair in the FRET network. Additional FRET pairs are chosen such that they maximize the product between the average maximum dynamic shift and the average absolute vectorial cross product between the candidate FRET pair and all previously selected FRET pairs added to the network. This ensures that selected FRET pairs span complementary distances that are likely to be useful for FRET-guided structural modeling.
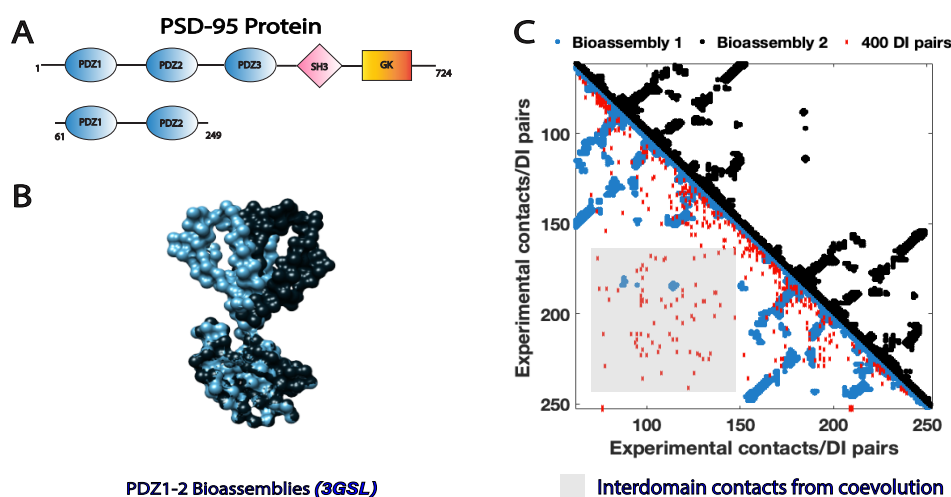
Figure 1. DCA uncovers the experimental contacts of PDZ1−2 tandem. (A) Domain representation of the 724-residue long full-length PSD-95 protein and the 196-residue long truncated tandem PDZ1−2 domains. (B) C-α representation of two different X-ray crystallography-derived forms of PDZ1−2 tandem (PDB ID: 3GSL a/b, BA1- blue, BA2- black) with an RMSD of 3.8 Å relative to one another. (C) Experimental contacts of BA1 (blue) and BA2 (black) of PDZ1−2 tandem are illustrated in a contact map with distance cut off 10 Å and the top 400 DI pairs (red) are overlaid on top of BA1's contacts with the gray box highlighting the interdomain residue pairs from coevolution.

(7) The selection of FRET pairs is terminated. In general, FRET pair selection can be extended to all possible FRET pairs, with the result being a ranked list of FRET pairs given the first chosen FRET pair as a "seed" for the network. Alternatively, FRET pair selection can be terminated given an uncertainty threshold within which structures resulting from modeling with the designed FRET network must fall relative to the target structures. The resulting FRET network of a chosen number of FRET pairs can be tested in several ways, such as simulating FRET observables with AVs and MD simulations in order to ensure the network can sufficiently reproduce the target states. Selection of FRET pairs in this work was terminated when the FRET network contained 10 pairs of labeling sites for direct comparison to the FRET network used for previous experiments.

Accessible Volume and Rigid Body Docking Simulations. Accessible volume (AV) simulations were performed using the *AVTraj* Python package, available at https://github.com/Fluorescence-Tools/avtraj. AV simulations of fluorophores are useful for calculating expected FRET observables from molecular dynamics simulation trajectories or individual biomolecular structural models and can further be used for FRET-guided modeling of biomolecular structural conformations.[17] For all AV simulations in FRET network design and rigid body docking, the AV1 (one-sphere fluorophore) model was used, with a linker length of 20.5 Å, linker width of 2.5 Å, and fluorophore radius of 5 Å. The C-α position of each residue was used as the linker attachment site, with residue side chains removed. Calculation of the FRET efficiency for each pair of AVs was performed by sampling 10000 distances between random points in the donor and acceptor AV clouds.

Rigid body docking (RBD) simulations were performed using the FRET Positioning and Screening (FPS) software package, available at https://www.mpc.hhu.de/software/fps.html. FPS provides a graphical user interface for performing AV simulations as well as RBD of biomolecular structural models such that simulated FRET observables from AVs best satisfy experimental or other distance constraints, thus producing a structure consistent with experimental observables.[17] For RBD, the PDZ1 and PDZ2 domains of the PDZ tandem were treated as separate structures, with the linker residues in the range 151−155 removed from both structures. Separate RBD was performed for the previously used FRET network and the FRET network from DCA-guided design. In each case, RBD was performed using all mean 10 distances restraints for one of the 4 target state clusters simultaneously. Thus, eight structures resulted from RBD, one for each target structure for each of the two FRET networks (Figure S2). For comparison of the resulting structures to the target structures, RMSD values were calculated using the CE-Align function of PyMol.[55]

Simulation of Single-Molecule FRET Data. Simulations of single-molecule data are done via Brownian Dynamics.[56,57] We assumed a 3D Gaussian distribution for the spatial intensity distribution of the observation volume (V_{Size} = 5 μm^3, with plane radii w_0 = 0.5 μm and axial radius z_0 = 2.24 μm. Each simulation had a total of N 50000000 photons, and a time step of 0.005 ms. We assumed a Gaussian Instrument response function with full width half-maximum (fwhm) of 0.3 ns and background scatter in the green channels of 0.5 kHz and 0.2 kHz in the red channels. Also, we consider a dark count rate of 0.2 kHz in the green and 0.1 kHz in the red channels.

We simulated each FRET variant following the pairing sites of OL, CL, Twisted, and Extended conformations as in Table S1. In each simulation, we assumed a population fractions 40% (OL), 30% (CL), 20% (Twisted), and 10% (Extended). These fractions correspond to an effective number of molecules in the observation volume (N_{fcs}) as 0.0012, 0.0009, 0.0006, and 0.0003, respectively, at any given time, thus ensuring single-molecule detection. The FRET distances and corresponding count rate for each channel (green and red) assumed a molecular brightness of the unquenched donor as 200 kcps and a fluorescence lifetime of 4 ns. We consider a constant rotational correlation time to 1 ns. Data are saved in Becker-Hickl SPC132 data format from which they are further analyzed as experimental data.

Analysis of Time-Resolved Fluorescence Decay Histograms. Analysis of fluorescence decay histograms from FRET experiments and simulated data with time-correlated single-
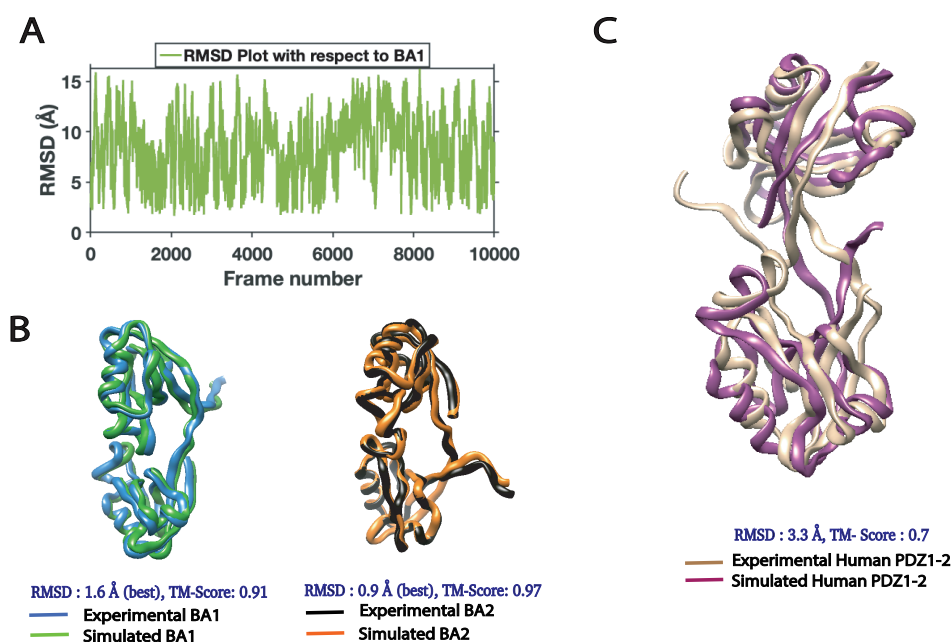
Figure 2. SBM+DCA captures the crystal structures of PDZ1−2. (A) RMSD of trajectory frames with respect to BA1. (B) Structures of the predominant frames in the simulation trajectory with the lowest RMSD with respect to BA1 and BA2 were aligned with the corresponding crystal structures. The hybrid model simulations captured the conformation of the human form of PDZ1−2 which was aligned with experimental structure (PDB ID: 3ZRT) as shown in (C).

photon counting (TCSPC) were performed as previously described[8] using the ChiSurf software package available at https://github.com/Fluorescence-Tools/chisurf. Fit models of experimental data consisted of states corresponding to three Gaussian-distributed interdye distances plus a donor-only fraction for each FRET variant. The donor-only fraction and fluorescence lifetime, as well as the Förster radius were fixed according to the values used in the previous study.[8] The population fractions for all three Gaussian-distributed states were set as global parameters among the globally fit fluorescence decay curves. For each of the twisted and extended states, previously used FRET variants were chosen for analysis based on meeting the criteria that the expected interdye distance for the representative structure deviated by at least 15% from all previously determined FRET-derived distances for a given variant, on the basis that fitting to the unexplained third state may have previously obscured evidence of these structures. 15% is well beyond the expected uncertainty in the FRET-derived interdye distances in confocal experiments.[14] For the Twisted state, variants E135-Y236 and S142-M159 were globally analyzed, while for the Extended state, variants Q107-Y236, D91-Y236, and M149-A230 were globally analyzed.

For simulated single-molecule FRET data, we considered a multiexponential form

$$F_{norm}(t) = \sum_{i}^{3 \text{ or } 4} x_i \exp\left(-\frac{t}{\tau_{DA,i}}\right) \tag{7}$$

with three or four different FRET quenched fluorescence lifetimes, each corresponding to a conformational state.

■ RESULTS AND DISCUSSION

SBM+DCA Recapitulates Conformations of PDZ1−2 Tandem of PSD-95. PSD-95 is a dynamic, multidomain protein that consists of five subdomains—three PDZ domains

in tandem, an SH3 domain and a Guanylate Kinase like domain which are involved in protein interactions at excitatory synapses[58] as shown in Figure 1A. The three PDZ domains bind to important synaptic proteins linked to synaptic activity.[59] In this study, we mainly focused on studying the conformational dynamics of PDZ1−2 tandem, which are targets for designing drugs for ischemic stroke.[60,61] The PDZ1−2 tandem protein is 196 residues in length with the PDZ domains consisting of 86 residues and separated by a flexible linker. The crystal structures of two states of PDZ1−2 tandem of PSD-95 derived from X-ray crystallography (PDB ID: 3GSL) Bioassembly 1 (BA1, color: blue) and Bioassembly 2 (BA2, color: black) of the organism *Rattus norvegicus* are displayed in Figure 1B. The PDZ tandem has about 34000 sequences, after filtering sequences with long gapped regions and reweighting sequences with more than 70% identity. We performed DCA on this family of sequences and identified coevolving sites that can be used as a proxy for contact maps for both BA1 and BA2 (see Figure 1C). The experimental contacts of BA1 and BA2 are highlighted in blue and black, and the top 400 DI pairs, or coevolving coupled sites, are overlaid in red on top of the BA1 contacts (C-α distance = 10 Å) as shown in the contact map in Figure 1C. In Figure 1C we can see that these DI pairs covered most of the intradomain regions, while the clusters present in between the two domains are hypothesized to represent the interdomain contacts (highlighted in the squares of Figure 1C) of the multiple functional conformations of PDZ1−2 and transition states between them. The states corresponding to these contacts can be further analyzed using the SBM+DCA methodology. This illustrates that the collection of sequences in the PDZ family have enough diversity and coevolutionary information to provide insights about the distinct structures and dynamics of these proteins. As a reference, we have also analyzed the single PDZ domain family and shown the top 90 DI pairs in Figure
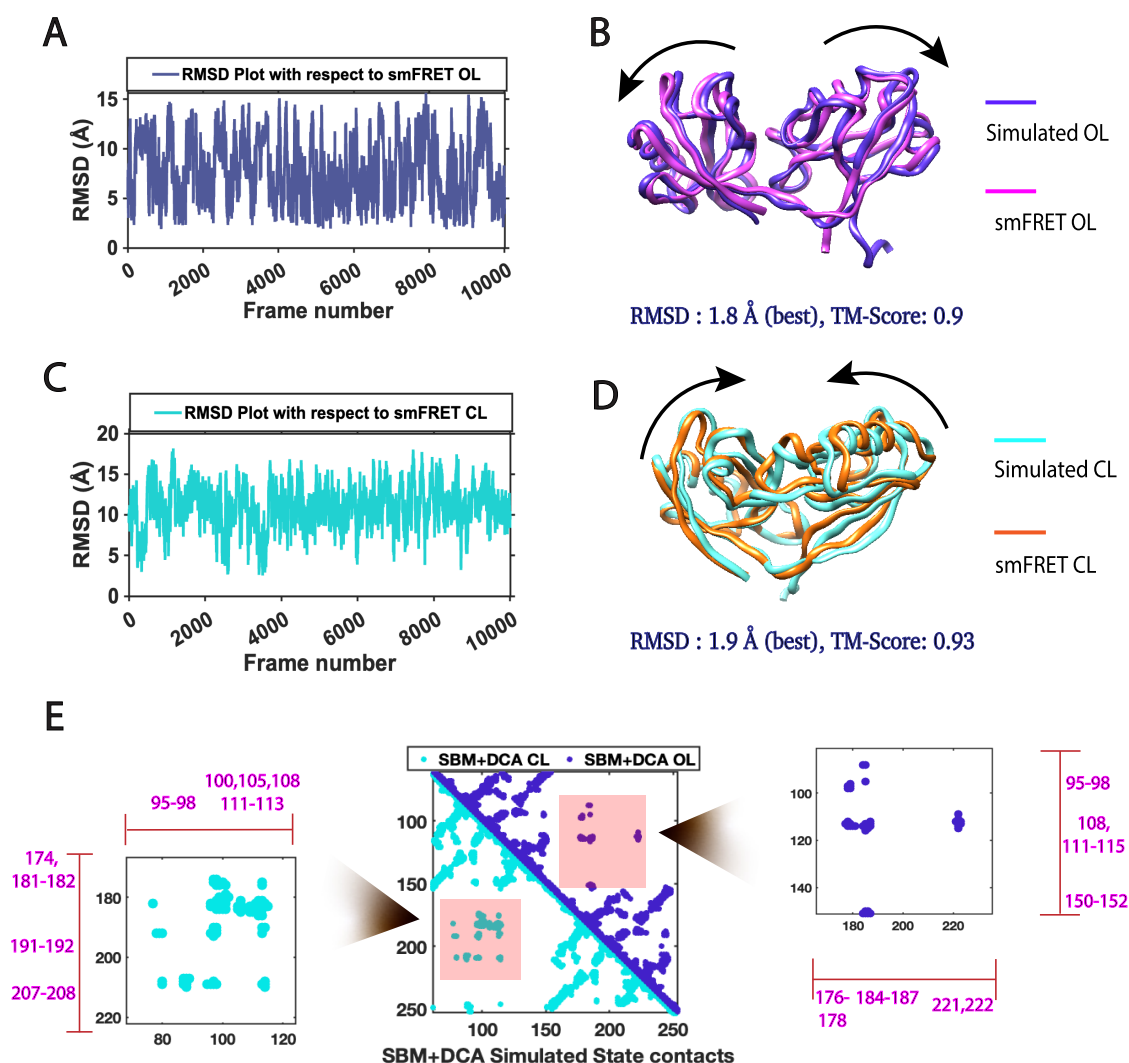
Figure 3. SBM+DCA visits states identified by smFRET experiments. (A) and (C) RMSD trajectory plots from SBM+DCA simulations with respect to experimental OL and CL states from smFRET experiments. (B) and (D) Predominant frames of OL state (purple) and CL state (green) from SBM+DCA simulations were aligned with their respective experimental structures (OL-pink, CL-orange) The RMSD values and TM-scores are indicated in the figures, respectively. (E) Interdomain contacts (highlighted in pink) of OL and CL states of PDZ1−2 tandem from SBM+DCA simulations matched with OL and CL states from smFRET experiments (C-α distance cut off:10 Å).

S3D. Although the single domain analysis has better predicted contacts, it misses important interdomain interactions.

To explore the conformational space of the PDZ family using coevolution, we applied the SBM+DCA model[47] (see details in the "Methods" section) to the PDZ1−2 tandem of PSD-95 protein. In this methodology we run C-α simulations, by using the structure parameters of the BA2 as the initial topology, and the top 400 coevolutionary couplings shown in Figure 1C were integrated into the model to drive the simulations to explore the conformational landscape of PDZ1−2 tandem. The root-mean-square deviation (RMSD) plot for the entire SBM+DCA trajectory with respect to BA1 in Figure 2A shows that the state PDZ1−2 BA1 was visited. We also used a template modeling score (TM score) to estimate topological differences between two protein structures. This TM-score gauges the global fold similarity between the structures by assigning stronger penalties for smaller distance errors than the larger distance errors between the structures and hence it is considered as an alternative metric to RMSD that we used in our study to compare and measure the folds of

two protein structures.[62] In our simulation trajectory, more than 20% of the structure frames had at most 3 Å RMSD with respect to BA1 and BA2 (Figure S4). We aligned the frames (with least RMSD and TM score) with the X-ray crystallographic PDZ1−2 BA1 and BA2 structures. The RMSDs and TM scores of the structures aligned with respect to the experimental BA1 were 1.6 Å and 0.91, and these were 0.9 Å and 0.97 with respect to BA2, as illustrated in Figure 2B. Although the initial structure used in the simulation was BA2, we can see that the DCA-predicted couplings were sampling the BA1 conformation accurately. We can see that the states BA1 and BA2 are not considerably different (RMSD 3.8 Å), but it is not trivial enough to ignore the difference between the states and call them alike. In addition, the SBM+DCA simulations also captured the conformation of the human form of PDZ1−2 tandem and the cluster centroid structure shown in Figure 4A was aligned with crystal structure (PDB ID: 3ZRT, here called the Extended conformation) as shown in Figure 2C, resulting in an RMSD and TM score of 3.3 Å and 0.7. These findings suggest that the combination of coevolu-
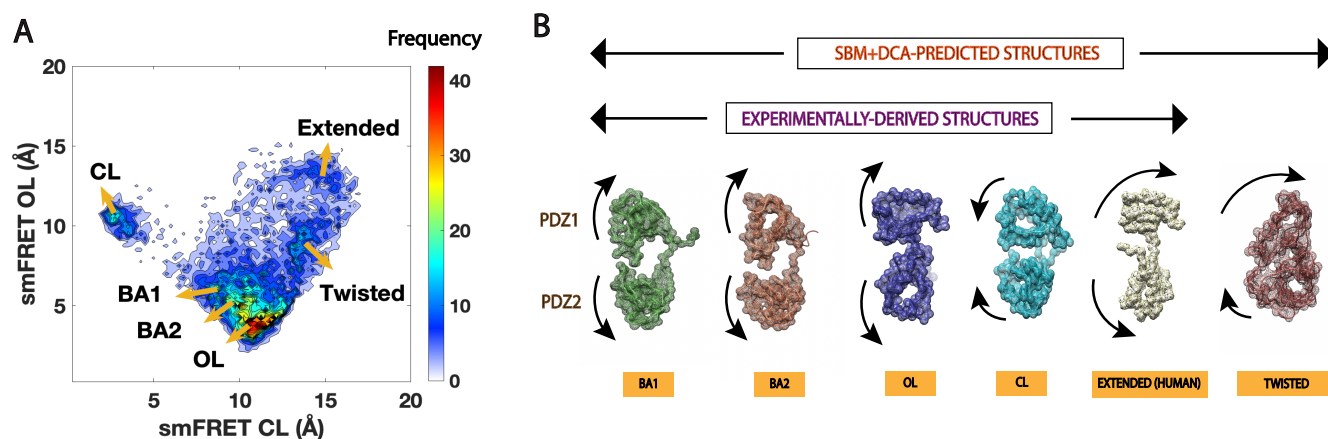
Figure 4. Elucidation of an unidentified twisted state of PDZ1−2 tandem from coevolution. (A) RMSD clustering of the SBM+DCA trajectory of PDZ1−2 tandem shows the clusters for six different states: Bioassembly 1 (BA1), Bioassembly 2 (BA2), open-like (OL), closed-like (CL), Twisted and Extended (Human) PDZ1−2 states with arrows pointing to the centroid of each state. (B) SBM+DCA methodology proposed structures (BA1- green, BA2- orange, OL- purple, CL- cyan, Extended (Human)- yellow, Twisted- red).

tionary information and coarse-grained models like SBMs can sample different conformations of PDZ1−2 tandem including those derived from different organisms with high accuracy.

SBM+DCA Uncovers States Identified by smFRET Experiments. In an earlier study, we showed that discrete molecular dynamics (DMD) simulations and single-molecule fluorescence resonance energy transfer (smFRET) studies on PDZ1−2 tandem identified two low energy states: an OL state and a CL state.[8] We examined if these sampled states were also captured by coevolution. We found that our SBM+DCA simulation trajectory, in fact, visited the OL and CL states. Figure 3A,C shows RMSD simulation trajectory plots with respect to these states. The structures of the principal frames with lowest RMSD values from the simulations relative to the aligned structures obtained from screening of simulation structures against smFRET-derived distance restraints are shown in Figure 3B,D. These simulated states had an RMSD of 1.8 and 1.9 Å with respect to smFRET OL and CL structures and TM-scores of 0.90 (OL) and 0.93 (CL) respectively. To better understand the spectrum of conformational states, we analyzed the differences in their interdomain contacts. The contact maps in Figure 3E show the unique set of interdomain contacts formed in each conformation (purple, contacts of OL state from DCA simulations; cyan, contacts of CL state from DCA simulations). Interestingly, we found that most of the interdomain contacts highlighted in this figure for both OL and CL conformations of PDZ1−2 were present in the regions similar to the interdomain contacts of the states predicted by smFRET experiments.[8] Among those contacts, 113−186, 113−178, 113−222, and 151−186 in the OL and 97−182, 113−183, 99−183, 99−184, and 114−208 in the CL showed perfect alignment with the efficiency interdomain contacts in the FRET-derived structures as shown in Table S2. Thus, the SBM+DCA protocol in predicting successfully predicts all the experimentally derived structures of PDZ1−2 tandem.

Coevolution Predicts an Unidentified Conformational State of the PDZ1−2 Tandem. Interestingly, when we performed RMSD clustering analysis with respect to the smFRET OL and CL states, we found that the SBM+DCA methodology predicted a unique cluster: an intermediate twisted PDZ1−2 conformation, with an RMSD of about 8 Å with respect to smFRET OL structure and 13.5 Å compared to

the smFRET CL structure as illustrated in Figure 4A. We selected one of the representatives from this most populated cluster to visualize the structure and we found that one of the PDZ domains was twisted in relation to the other domain as shown in the last structure on the right (in red) in Figure 4B. From the simulation trajectory, we found that the twisted conformation was an intermediate state taken by PDZ1−2 tandem when transitioning between the OL and CL states. Additionally, the interdomain contacts of this twisted state are illustrated in the Figure S3A and this state had a distinct set of interdomain contacts (highlighted in green boxes) except for a few regions that matched the OL and CL states. There were six DI pairs that exactly matched the interdomain contacts of the twisted state and 25 DI pairs in close proximity to them (brown boxes in the Figure S3B) that may have driven the simulations toward this twisted state. The SBM+DCA simulation trajectory revealed five different experimentally determined states of PDZ1−2 tandem: BA1, BA2, OL, CL and Extended (human) as shown in Figure 4A,B. Each of these states is represented as a cluster in Figure 4A, and the frequency is the count of structure frames in the trajectory for a specific RMSD coordinate. Out of these states, OL and CL states, obtained from FRET-derived restraints, have about a 10 Å RMSD difference between each other as shown in Figure 4A,B. BA1 And BA2, obtained from X-ray crystallography, fall within the OL cluster as seen in Figure 4A. Apart from the five states, the simulations proposed a sixth state, the twisted state, that may be a functional intermediate state but that requires further experimental validation. Although the cluster near 13 Å RMSD relative to OL and 15 Å relative to CL is not a well-separated basin, the centroid corresponds closely to the extended conformation of human form PDZ1−2. This extended conformation exists within a large, dispersed cluster population that is not driven by interdomain interactions, as almost no interdomain contacts exist (Figure S3B). As a control, we ran simulations without any enrichment from coevolution (Figure S5A) and with the same number of integrated pairs used before but chosen randomly (Figure S5B). As expected, we observe no further exploration of relevant conformations in this control simulation. Agreement between the SBM+DCA structures and those derived from X-ray crystallography but previously not identified via smFRET served as a motivation to synergize SBM+DCA with smFRET
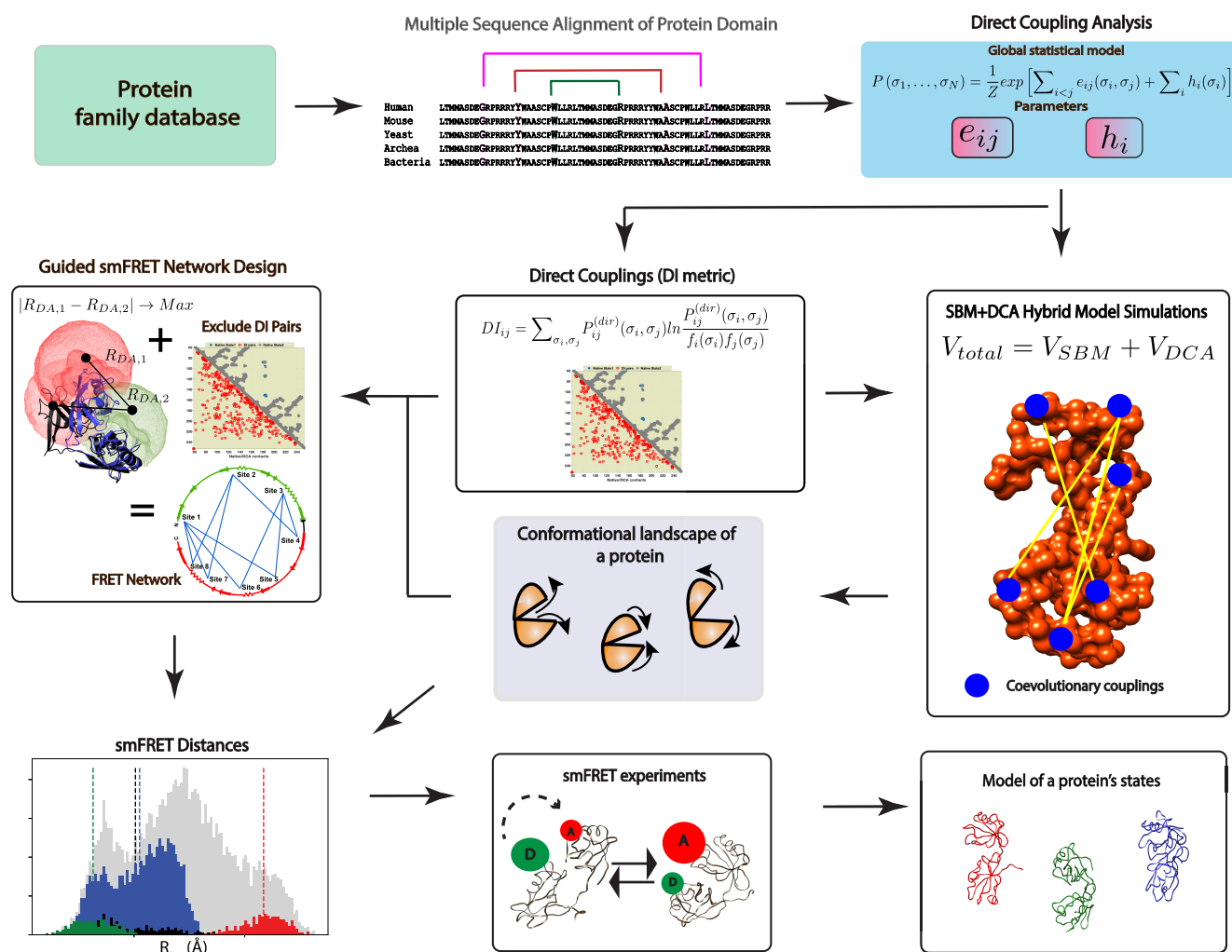
Figure 5. DCA + smFRET Workflow: The MSA of the sequences of PDZ family from the Pfam database is extracted, and DCA is applied to the MSA to calculate the parameters e_{ij} and h_i. These parameters are then used to score and identify strong coevolutionary couplings, which are incorporated into the SBM energy function. The conformational transitions of a protein from these SBM+DCA simulations are used to predict the residue positions that could be involved in the dynamics of a protein. Further, representatives for each major basin from simulations, as well as DCA-predicted contacts were used as inputs for the design of a FRET network targeted at maximally resolving four distinct conformations of the PDZ tandem. The FRET network was optimized for resolvability of structures via modeling with FRET observables while avoiding perturbations to residue−residue interaction sites predicted via DCA. These labeling sites were tested *in silico*, using simulated FRET observables for all SBM+DCA structures, to verify that predicted experimental observables could reproduce the target structures. Finally, while not performed for this study, variants of the PDZ tandem could be produced for each FRET pair in the FRET network and measured experimentally to probe for the predicted conformations.

to further probe the PDZ1−2 tandem twisted and extended (human) states.

Coevolution and FRET-Network Design Using Predicted States. The ability of SBM+DCA to recapitulate and predict novel conformational states of proteins makes it well-suited to use in designing experiments that might confirm these states. In particular, SBM+DCA provides two crucial outputs for the design of experiments which rely on targeted labeling of proteins for resolution of distinct conformations, such as in single-molecule FRET experiments that require attachment of fluorophore labels to the protein surface.[11] First, DI pairs provide information about amino acid positions which, based on evolutionary pressures, may be involved in the adoption of a particular state and which may not otherwise be documented in the literature. Such amino acid positions should be avoided in labeling to avoid perturbing the native behaviors of proteins. Second, the predicted states provide

targets for informed selection of labeling sites. Labeling sites are chosen such that they maximize the ability of changes in the state of the labeled molecule among the target states to be observed via experimental observables. Thus, we propose a simple workflow for design of FRET networks or sets of labeling site pairs which are measured in separate FRET experiments, for mutually resolving multiple target states simultaneously. Data from multiple experiments with different labeling sites in a FRET network can be analyzed together to model the limiting conformational states of dynamic proteins.[8,17] The designed protocol is summarized in Figure 5. First, a protein of interest is identified (in this case PDZ1−2 tandem), and a database of sequences for evolutionarily related proteins is compiled. Next, these sequences are aligned along the regions of interest prior to DCA. The direct couplings (pairs with high DI) resulting from DCA are then used to impart additional potentials on the amino acids involved in DI

pairs for SBM+DCA simulations. These potentials drive the protein toward conformations that satisfy the contacts predicted by DCA. Thus, SBM+DCA is well-suited to obtaining representative structures for limiting states like those modeled via FRET networks while cutting down on the computational cost associated with transitions through transient intermediate states. Next, clustering or similar analysis techniques are used to select centroid or representative structures for the conformational states of interest that will serve as target states for FRET network design. Then, the DI and target structures from DCA are combined with simulations of FRET observables to design a FRET network to mutually resolve the target states. FRET network design proceeds as described in the "Methods" section.

To test the ability of the designed FRET network to model the distinct protein conformations, we performed accessible volume (AV) simulations and calculated predicted FRET observables for all structures from SBM+DCA simulations. The simulated FRET observables are then used as restraints for rigid body docking simulations to model the distinct conformations corresponding to the FRET observables for the target state basins from SBM+DCA simulations. Here, we defined these basins via RMSD clustering of simulation structures against the target states. Finally, if positive results are obtained via modeling with simulated observables, then the labeled protein samples for the designed FRET network can be produced, measured, and analyzed for cross-validation of predicted structures and quantification of the dynamics of conformational exchange for the protein.

We previously integrated results from smFRET experiments and discrete molecular dynamics (DMD) simulations to identify the distinct OL and CL conformations of the PDZ1−2 tandem.[8] Further, we determined that previous conformations identified via NMR and TIRF-based smFRET were compatible with apparent dynamic averaging of the OL and CL states. While the conformations from crystal structures (3GSLa,b) had associated interdomain distances slightly beyond those corresponding to the OL and CL structures, the simulated interdye distances were similar to those for the OL state. Further, the RMSD analysis in Figure 4A indicates that these structures are within the OL ensemble. In modeling the OL and CL states from smFRET data, a three-state global model was used for determination of interdye distances for all 10 used FRET pairs simultaneously. Two of the sets of 10 distance distributions resulted in the OL (for the dominant population fraction) and CL (for the middle population fraction) states, respectively, from screening of FRET-derived distances against expected interdye distances for all structures from DMD based on accessible volume (AV) simulations of the fluorophores (examples shown for OL and CL states in Figure S6). These two major populations agreed with the major populations from DMD simulations. The third set of distances did not produce a model that satisfied all distance restraints simultaneously. Additionally, the DMD simulations produced a significant third population (though less populated than the OL and CL basins) of structures with interdomain distances beyond those for the OL and CL states, but for which predicted FRET observables had significant overlap with either one or both of the OL and CL states for several FRET pairs, presenting a barrier to resolving these structures via the used FRET pairs. Furthermore, distances obtained for the third set of distance restraints were not globally consistent with any of the three major populations from DMD, while the OL and

CL distances were consistent with the respective basins. Meanwhile, the predicted interdye distances for some FRET pairs associated with the Twisted and Extended states identified with DCA are distinct from those assigned to the CL and OL states. Thus, we identified five of the previously used FRET pairs for which either the Twisted (two pairs, E135-Y236 and S142-M159) or Extended (three pairs, Q107-Y236, D91-Y236, and M149-A230) state distances differed significantly from those for OL and CL, based on the lowest relative percent difference between the predicted distance and all 3 previously identified distance populations, and reanalyzed the fluorescence decay histograms for these samples. This analysis was performed using a three-state model with global population fractions for each subset of FRET pairs. The distances resulting from this analysis are provide in Table 1. The results of these fits were compatible with the presence of Twisted state as a minor population for S142-M159, while for E135-Y236 the expected Twisted state distance was not apparent. M149-A230 exhibited a major population compatible with the extended state. Overall, the results of these fits suggest that these states may exist, but that the presence of other states with similar interdye distance distributions in the 10-variant global modeling may have obscured them. While this analysis does not confirm these states, the expected overlap among simulated FRET distances for the DMD basins suggests this may result from the FRET network not being optimally designed for resolving these states, even if they are present.

DCA-Guided Design of FRET Network. We used the previously described protocol to generate a FRET network of 10 distinct FRET pairs that would mutually resolve four target conformations of the PDZ1−2 tandem (OL, CL, Twisted, and Extended). The four target representative conformations were chosen because they represent conformational basins that are well separated from each other in Figure 4. Thus, the goal of the DCA designed FRET network is to maximally resolve these distinct basins. The proximity of the 3GSLa (BA1) and 3GSLb (BA2) structures to the CL basin suggest that they may appear as part of the OL basin when modeling with experimental constraints because experimentally derived constraints are subject to dynamic averaging when they are in fast exchange. The Extended and Twisted states were included as they are well-separated from the OL and CL basins in SBM+DCA simulations and further were not found in the previous experimental work. We chose to terminate the FRET network at 10 FRET pairs so that we could directly compare the outputs of the design protocol against a previously used FRET network for the same protein.[6] Representative structures for the four target states from DCA simulations used in FRET network design are shown in Figure S7. First, residues appearing in the top 1000 ranked DI pairs or with contacts apparent in the target states were removed from the list of candidate labeling sites for FRET pair selection. Additionally, candidate sites appearing in the interdomain linker between the domains were excluded from consideration. Next, accessible volume (AV) simulations of the fluorophores were performed for all remaining labeling site candidates (Figure S6), and AVs were sampled to compute expected interdye distances for all possible pairwise combinations of labeling sites. The final FRET network was generated by first selecting the candidate FRET pair which maximized the average difference in FRET efficiency among all four target states. Additional labeling sites were selected by maximizing the average change in FRET efficiency times the average vectorial

Table 1. Predicted and Refitted Mean Interdye Distances of FRET Variants That May Report on Experimentally Unresolved PDZ Tandem Conformations[a]

	predicted ⟨R_{DA}⟩ values		site 1	site 2	refits				derived from the global fluorescence lifetime analysis			
	⟨$R_{DA,sim,Twist}$⟩ (Å)	⟨$R_{DA,sim,Extend}$⟩ (Å)			⟨$R_{DA,new\_exp,1}$⟩ (Å)	⟨$R_{DA,new\_exp,2}$⟩ (Å)	⟨$R_{DA,new\_exp,3}$⟩ (Å)	χ_r^2	⟨$R_{DA,exp,1}$⟩ (Å)	⟨$R_{DA,exp,2}$⟩ (Å)	⟨$R_{DA,exp,3}$⟩ (Å)	χ_r^2
	44.1		E135	Y236	59.2 ± 3.0	70.0 ± 1.4	36.2 ± 7.4	1.32	74.8 ± 1.8	57.7 ± 1.7	32.6 ± 0.8	1.23
	35.2		S142	M159	48.8 ± 3.4	61.3 ± 2.5	35.1 ± 2.5	1.29	60.3 ± 1.1	53.1 ± 0.9	47.8 ± 0.8	1.08
fraction (%)					**49.5**	**34.2**	16.3		43.5	32.4	24.1	
		72.8	Q107	Y236	62.5 ± 1.9	48.2 ± 3.9	31.9 ± 2.2	1.22	37.8 ± 0.9	57.8 ± 1.4	15.7 ± 1.1	1.10
		46.0	D91	Y236	38.4 ± 0.8	55.4 ± 1.7	33.3 ± 7.0	1.27	54.3 ± 2.0	31.8 ± 1.1	29.7 ± 1.1	1.08
		41.8	M149	A230	51.8 ± 0.6	40.6 ± 2.0	60.3 ± 2.4	1.21	49.6 ± 0.5	59.2 ± 0.6	31.8 ± 0.3	1.08
fraction (%)					**45.5**	**39.5**	15.0		43.5	32.4	24.1	

[a]From prior work, five FRET variants were identified for which ensemble or dynamic averaging of distances derived from the global fluorescence lifetime analysis of ten previously designed variants (italic font, right-hand columns) would not likely recapitulate predicted ⟨R_{DA}⟩ values (normal font, left-hand columns) for either the predicted Twisted or Extended state. These variants were reanalyzed globally among those identified for Twisted (E135-Y236 and S142, M159) and for Extended (Q107-Y236, D91-Y236, and M149-A230). The resulting fits (bold font, center columns) were of similar overall fit quality to the previous analysis, and although population fractions were like those derived from the previous analysis, assignments of distances to either the major, middle, or minor population fractions were not conserved. In the previous work, the third population ⟨$R_{DA,exp,3}$⟩ could not be assigned to a particular structure, suggesting these distances may correspond to averaging among heterogeneous structures rather than a single state. For some fits (S142-M159, D91-Y236, and M149-A230), the ⟨$R_{DA,exp,1}$⟩ and ⟨$R_{DA,exp,2}$⟩ distances were recapitulated from the new analysis, while an additional distance was observed similar to the Twisted and Extended states were observed in experiments but averaged out, but it suggests that such a determination would require a FRET network designed with these potential conformations in mind. Distance bounds represent the 95% confidence interval for the mean interdye distances from fluorescence decay fits, determined via the F-test for the ratio of χ_r^2 under variation of each model parameter as in previous work.[8]

cross product between the candidate FRET pair's interdye distance vector and those for all previously selected FRET pairs. This additional cross product factor was used to minimize the redundant information provided by the addition of new FRET pairs considering previously selected sites. This procedure was repeated until the designed FRET network consisted of 10 distinct FRET pairs. Both the previously used and newly designed FRET networks are represented in Figure 6 and Tables 2 and S3.

To evaluate the performance of the FRET network resulting from DCA-guided design relative to the previously used FRET network in resolving the four target states, we simulated accessible volumes (AV) from which we computed FRET observables for all FRET pairs from both FRET networks for all structures from DCA-guided MD simulations. All structures
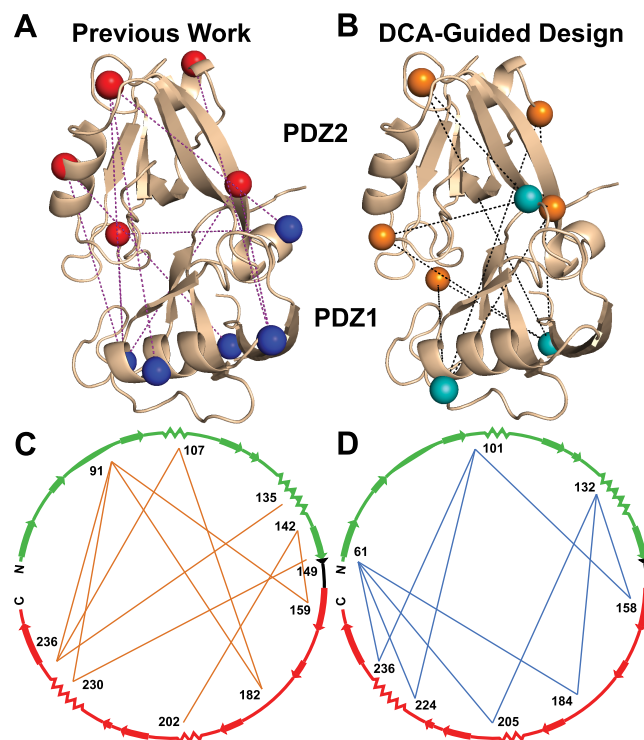

Figure 6. Graphical representations of FRET networks from previous work and DCA-guided design. (A) Set of FRET labeling sites for ten FRET variants used in previous work for determination of two dynamically exchanging conformations of the PDZ tandem (open-like and closed-like states). The ten pairs of labeling sites constitute a FRET network which was analyzed globally. (B) FRET network resulting from DCA-guided empirical design, in which structures from DCA as well as DCA-predicted contacts provided inputs for selection of labeling sites. All sites were screened for labeling sites that provided the best mutual contrast among all four target structures. Labeling sites with high DI based on DCA analysis were avoided in labeling site selection in order to avoid perturbing native contacts. Labeling sites were chosen iteratively based on highest average contrast among all states while maximizing the magnitude of the cross product, computed using the vectors between labeling sites, between additional and previously chosen site pairs. Additionally, labeling sites within two or fewer residues of previously chosen sites were merged into the previous sites. (C, D) Alternative representations from networks from (C) previous work and (D) DCA-guided FRET network design. Circles represent linear sequences of the PDZ1−2 tandem, with PDZ1 colored green, PDZ2 colored red, and the interdomain linker in black.

Table 2. Interdye Distances from Single-Molecule Simulations[a]

Site 1	Site 2	CL $R_{DA,Bur}$ (Å) (% error)	OL $R_{DA,Bur}$ (Å) (% error)	Twisted $R_{DA,Bur}$ (Å) (% error)	Extended $R_{DA,Bur}$ (Å) (% error)
135	236	53.6 (3.5%)	54.3 (2.2%)	46.0 (3.6%)	59.2 (4.7%)
107	236	42.1 (0.1%)	42.1 (0.1%)	35.7 (2.3%)	57.9 (<0.1%)
91	159	32.6 (<0.1%)	38.4 (1.4%)	38.3 (3.7%)	34.0 (3.9%)
91	182	30.7 (3.9%)	23.9 (13.0%)	30.7 (3.9%)	34.2 (7.9%)
142	159	46.3 (1.6%)	48.0 (2.5%)	38.8 (0.1%)	50.1 (1.9%)
107	182	23.9 (4.9%)	34.4 (2.9%)	34.5 (2.6%)	40.3 (1.8%)
91	236	47.1 (1.7%)	43.6 (3.3%)	47.4 (6.8%)	47.6 (1.8%)
91	230	42.8 (<0.1%)	39.7 (0.1%)	35.6 (1.1%)	39.7 (1.8%)
149	230	39.4 (0.1%)	39.4 (1.8%)	36.9 (1.9%)	39.4 (0.1%)
142	202	45.6 (1.9%)	57.8 (2.6%)	58.4 (10.0%)	53.6 (11.0%)
61	236	50.7 (0.5%)	50.8 (0.4%)	50.7 (0.2%)	46.2 (0.7%)
61	205	37.5 (1.3%)	37.5 (1.8%)	39.6 (0.5%)	37.5 (0.5%)
101	224	25.9 (6.9%)	26.0 (3.7%)	29.1 (3.8%)	53.7 (0.2%)
101	158	42.5 (0.5%)	38.2 (3.7%)	35.2 (9.0%)	35.2 (9.4%)
61	224	40.1 (0.6%)	40.1 (1.1%)	40.1 (0.8%)	34.4 (6.4%)
61	183	28.1 (0.4%)	28.1 (5.1%)	28.1 (4.4%)	32.9 (24%)
149	158	34.5 (0.7%)	31.3 (0.7%)	31.3 (1.7%)	31.3 (2.0%)
101	236	40.5 (3.6%)	40.5 (0.6%)	40.5 (19.0%)	54.78 (11.0%)
132	205	39.9 (0.7%)	57.8 (0.6%)	57.8 (1.9%)	57.8 (0.4%)
132	183	40.4 (0.6%)	38.4 (2.1%)	38.5 (3.9%)	44.3 (0.3%)

[a]The distances were computed from the determined fluorescence lifetimes in Table S1 using

$$R_{DA} = R_0 \left(\frac{1}{1 - \frac{\tau_{DA}}{\tau_{D(0)}}} - 1 \right)^{1/6} \tag{8}$$

where the fluorescence lifetime of the donor in the absence of acceptor ($\tau_{D(0)}$) was set to 4 ns, and the Förster radius (R_0) was 52 Å. The error in accuracy relative to the expected distance from the target structure (Table S2) are shown in parentheses. Rigid body simulations used these distances from either the old (normal text, top half) or new (bold text, bottom half) DCA-guided FRET network simultaneously as distance restraints to test how well the obtained distances could recreate the target structure.

from DCA-guided MD were clustered into either the CL, OL, Twisted, or Extended state based on a 6 Å RMSD cutoff with respect to the target structure used in FRET network design. The resulting clusters are represented in histograms of AV-simulated interdye distances for all structures in Figure S10. The mean and standard deviation of the interdye distance distributions were calculated for each cluster (Tables S3 and S1 and Figure S10) and used for further analysis. Here, we use the mean distance corresponding to each cluster for analysis because the exact interdye distance distributions corresponding to individual conformations are experimentally inaccessible even for single-molecule techniques. This approach is analogous to using interdye distances derived from time-resolved FRET experiments, as is presented below, in which the distances correspond to the means of limiting state distance distributions and the widths of these distributions result from intrinsic fluorophore dynamics and extremely fast exchange processes. Because the design protocol discussed here seeks to maximize apparent contrast between target conformations, it is also applicable to techniques more prone to ensemble and time averaging, such as ensemble, intensity-based FRET. However, the distances derived from experimental techniques likely correspond to averages among the underlying conformations rather than to individual limiting states. While the widths of the individual clusters differed between the FRET networks, the average widths of the entire distance distributions were similar for both networks. However, the standard deviations among the means of the distributions were increased, on average, for the network from DCA-guided design (4.6 Å) when compared to the previously

used network (3.8 Å), indicating that the observed mean interdye distances for each cluster are more separated and thus provide greater contrast between the states as observed by FRET. The spread about the mean of the entire distance distribution was also increased for simulated distances for the newly designed network (Figures S8 and S9). We additionally used the sets of mean cluster R_{DA} values as distance restraints for rigid body docking (RBD) simulations to see how well the obtained distances reproduce the target structures, as would be performed in experiment. The resulting structures for each state are shown in Figure S2. To evaluate the accuracy of each structure, we calculated the RMSD for each structure resulting from RBD simulations with the target structure as reference. For the CL, OL, and Extended states, the FRET network from DCA-guided design outperformed the previously used network. The old network reproduced the Twisted state with slightly higher accuracy (3.3 Å vs 4.4 Å). Overall, the mean RMSD of resulting structures from the new network was more than halved when compared to the old network (7.8 Å for the old vs 3.5 Å for the new).

To evaluate how time-resolved FRET data can be used to generate structural models of the CL, OL, twisted and extended conformers, we simulated single-molecule FRET data considering a heterogeneous mixture of these conformations with population fractions of 40%, 30%, 20%, and 10%, respectively (see the "Methods" section). Single-molecule events were then selected by burst analysis, and the photon arrival times corresponding to the donor fluorescence were binned into histograms to build time-resolved fluorescence decays for all the FRET variants in the old and new DCA-
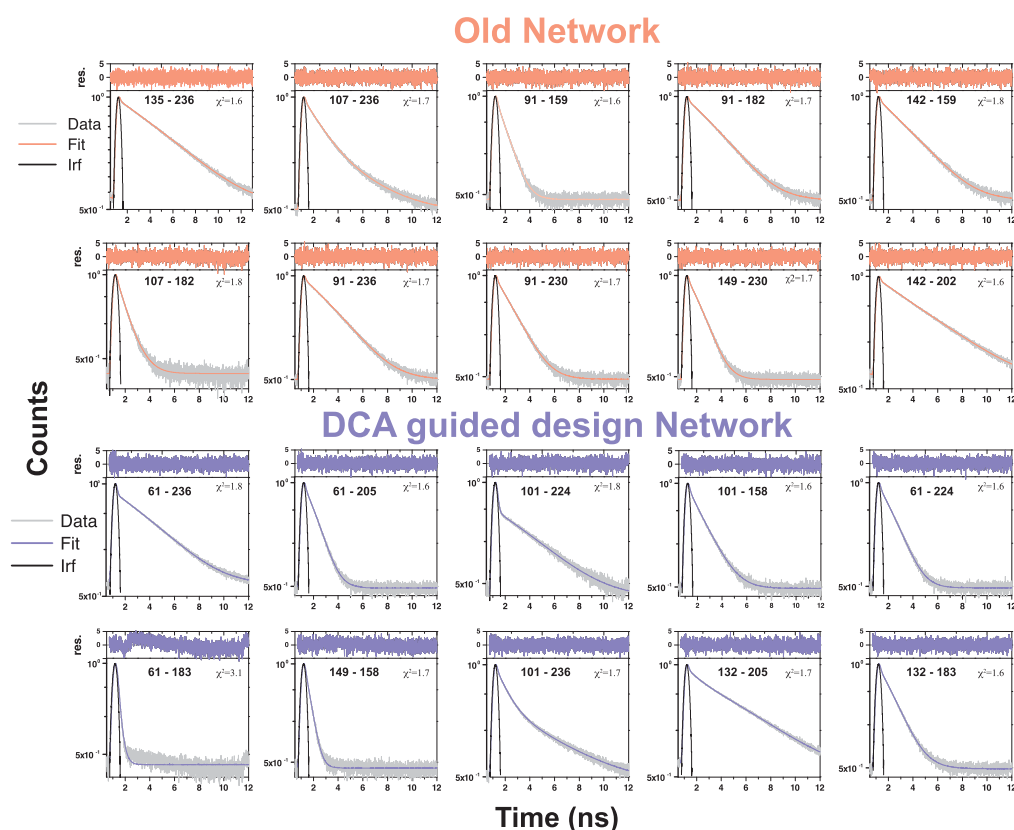
Figure 7. Time-resolved fluorescence decays for simulated single-molecule FRET data. Time-resolved decays for the old (salmon) and DCA-guided FRET network variants (blue) are shown. The instrument response function is shown in black, and the donor-only fluorescence decay is shown in gray. Raw histogram data are shown as points, with fits overlaid as colored lines. FRET quenches donor fluorescence, which reduces the lifetime and introduces curvature into fluorescence decay. The presence of more than one underlying state with different FRET efficiencies results in multiexponential fluorescence decays. A four-state model provided the best global fit for the 10 FRET variants in each network. Derived distances associated with the four states are found in Table 2. Details of the simulated data and fit model can be found in the "Methods" section.
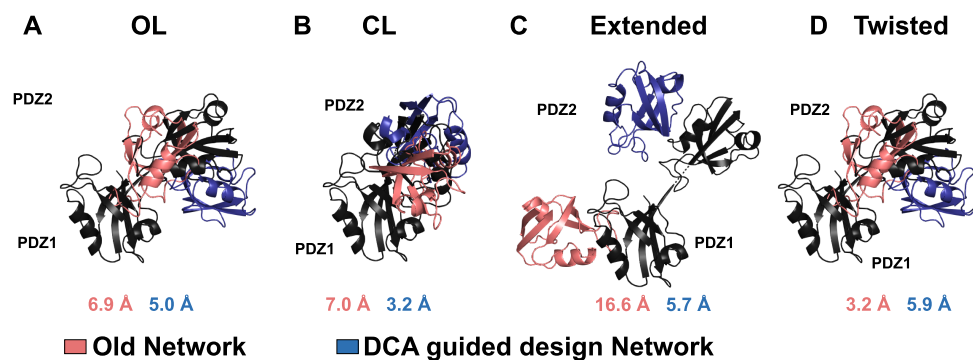

Figure 8. Structures resulting from rigid body docking of PDZ2 to PDZ1 based on simulated single-molecule experiments. Structures from distances in Table 2 corresponding to both the FRET network from the previous study (deep salmon) and the DCA-guided FRET network (deep blue) are shown for the OL (A), CL (B), Extended (C), and Twisted (D) conformations. Structures in black represent the target states from DCA-guided simulations that were used for FRET network design. RMSD values for each structure resulting from rigid body docking are shown below each set of structures. RMSD values were calculated using C-α positions with the CE-Align function in PyMol. The DCA-guided FRET network shows a significant improvement in using FRET-derived distance restraints to resolve four distinct target structures.

guided FRET networks (Figure 7). The time-resolved fluorescence decays for each FRET network were processed exactly as experimental data, for which we globally fit the FRET induced decays to determine the number of states and the corresponding limiting FRET distances. This analysis showed that using a four-state model best fit the FRET network in each case with average χ^2 values of 1.76 for the old network and 1.88 for the DCA-guided FRET network. When using a three-state model, χ^2 increased to 1.9 and 2.1, respectively. Moreover, 92.7% of the recovered FRET distances for both the old and DCA-guided networks were within the expected 7% accuracy of the target distance (Table 2). Only 7.5% of distances showed a larger deviation.

Next, we used the distances and associated uncertainties from the time-resolved decays ($R_{DA,Bur}$, Table 2) as distance restraints for rigid body docking (RBD) simulations as

previously done with the clustered DCA structures (Figure 8). We used the RMSD as a measure of the accuracy in resolving a structural model for each conformational state. Overall, the old network showed an average RMSD of 8.4 Å for all structures, while the DCA-guided FRET network design improved the average RMSD to 5.0 Å. Even when both networks are meant to represent the same structural models, the DCA-guided FRET network is better suited for resolving all four target states simultaneously. This is independent of the accuracy of the distances but rather specific to the selection of the site for the FRET labeling.

CONCLUSIONS

In this study, we combined coarse-grained protein models, coevolutionary information, and simulated and published smFRET data to explore the conformational landscape of the PDZ1−2 tandem and to guide smFRET experimental design. We identified at least six different states of the PDZ1−2 tandem selected by evolution since DCA was able to efficiently capture the residue contacts of PDZ1−2 tandem that provide information about its dynamics and the unique set of interdomain contacts that stabilize these states from different organisms. The SBM+DCA methodology not only captures the compact states with sufficient interdomain contact information but also predicts a potential transient state with no interdomain contacts, found previously for the PDZ1−2 tandem in humans. Of the six states identified, Twisted and Extended (human) states that were not observed in the previous smFRET study were found to be compatible with results from reanalysis of a subset of the FRET variants from the prior study, which previously resulted in an unexplained third set of interdye distance restraints that could not be explained by a single structural model. Thus, SBM+DCA was able to provide structural models that may resolve the conformations which previously gave rise to the unexplained distance restraints. However, the previously available FRET distances when used as restraints are not sufficient to fully model the predicted conformations. This, combined with the similarity in predicted interdye distances for different basins from previous DMD simulations, suggests that the previous FRET network may not be optimal for resolving all these states simultaneously. These observations have implications in designing smFRET networks, for which DCA provides two key pieces of information. First, SBM+DCA provides target structures of interest, driven by evolutionarily predicted contacts. We combine these resulting structures with techniques for simulating FRET observables to select FRET pair labeling sites which maximize contrast between the target structures. We also note that the use of a one-bead model SBM +DCA simulations accelerates the inclusion of DCA-based restraints. Thus, reducing the computation time associated with obtaining these structures is of interest, whereas traditional all-atom MD simulations require greater computational costs and explore the biomolecule's conformational landscape thoroughly. However, it is primarily the limiting conformations that are of interest for modeling via FRET. Second, the degree of coevolution and conservation in a protein family identified via DCA indicate the optimal residue positions and amino acids that may be important to the native conformations and the function of the protein of interest. This knowledge provides information about amino acids that should be avoided in labeling site selection to avoid perturbing the native conformations of the protein through attachment of the fluorophores.

To validate our approach, we simulated FRET induced fluorescence decays that compare two FRET networks even when both networks provide distances from the structural models. However, due to the sparse information that can be collected from FRET and optimized network is required for recovering structural models that are consistent with the target structures. By either clustering DCA simulated structures or simulating smFRET experimental data for the target structures, we found a significant improvement in the rigid body docked models that take into consideration solely the FRET derived distances for the four conformational states. Thus, the synergy between DCA and smFRET studies in guiding FRET network design paves a way to automation, to obtain the positions favorable for smFRET dye-coupling for several protein systems which can be beneficial to the smFRET community at large.

ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.2c06720.

> FRET network design workflow, structures from rigid body docking of PDZ2 to PDZ1, contact maps of PDZ12-tandem, correlation of simulation trajectory frames of PDZ1−2 tandem with respect to the bioassemblies, control simulations trajectories of PDZ1−2 tandem, example of fluorophore accessible volumes, target structures for FRET network design, histograms of mean distances for each target state, spread in R_{DA} for each simulated FRET pair, distance distributions from DCA-guided simulations, global fit of the four state models, comparison of the interdomain (ID) contacts of PDZ1−2 tandem states, summary statistics for AV-simulated R_{DA} distributions (PDF)

AUTHOR INFORMATION

Corresponding Authors

Faruck Morcos − *Departments of Biological Sciences and Bioengineering, University of Texas at Dallas, Richardson, Texas 75080, United States; Center for Systems Biology, University of Texas at Dallas, Richardson, Texas 75080, United States;* ⓞ orcid.org/0000-0001-6208-1561; Email: faruckm@utdallas.edu

Hugo Sanabria − *Department of Physics and Astronomy, Clemson University, Clemson, South Carolina 29634, United States;* ⓞ orcid.org/0000-0001-7068-6827; Email: hsanabr@clemson.edu

Authors

Aishwarya Krishnamohan − *Departments of Biological Sciences and Bioengineering, University of Texas at Dallas, Richardson, Texas 75080, United States*

George L. Hamilton − *Department of Physics and Astronomy, Clemson University, Clemson, South Carolina 29634, United States*

Rajen Goutam − *Department of Physics and Astronomy, Clemson University, Clemson, South Carolina 29634, United States*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jpcb.2c06720

Author Contributions

‖A.K. and G.H. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Medina, E. R.; Latham, D. R.; Sanabria, H. Unraveling Protein's Structural Dynamics: From Configurational Dynamics to Ensemble Switching Guides Functional Mesoscale Assemblies. *Curr. Opin Struct Biol.* **2021**, *66*, 129−138.

(2) Kasahara, K.; Terazawa, H.; Takahashi, T.; Higo, J. Studies on Molecular Dynamics of Intrinsically Disordered Proteins and Their Fuzzy Complexes: A Mini-Review. *Comput. Struct Biotechnol J.* **2019**, *17*, 712−720.

(3) Adamski, W.; Salvi, N.; Maurin, D.; Magnat, J.; Milles, S.; Jensen, M. R.; Abyzov, A.; Moreau, C. J.; Blackledge, M. A Unified Description of Intrinsically Disordered Protein Dynamics under Physiological Conditions Using NMR Spectroscopy. *J. Am. Chem. Soc.* **2019**, *141* (44), 17817−17829.

(4) Kodera, N.; Noshiro, D.; Dora, S. K.; Mori, T.; Habchi, J.; Blocquel, D.; Gruet, A.; Dosnon, M.; Salladini, E.; Bignon, C.; et al. Structural and Dynamics Analysis of Intrinsically Disordered Proteins by High-Speed Atomic Force Microscopy. *Nat. Nanotechnol* **2021**, *16* (2), 181−189.

(5) Zhang, J.; Lewis, S. M.; Kuhlman, B.; Lee, A. L. Supertertiary Structure of the MAGUK Core from PSD-95. *Structure* **2013**, *21* (3), 402−413.

(6) McCann, J. J.; Zheng, L.; Rohrbeck, D.; Felekyan, S.; Kühnemuth, R.; Sutton, R. B.; Seidel, C. A. M.; Bowen, M. E. Supertertiary Structure of the Synaptic MAGuK Scaffold Proteins Is Conserved. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (39), 15775−15780.

(7) Tompa, P. On the Supertertiary Structure of Proteins. *Nat. Chem. Biol.* **2012**, *8* (7), 597−600.

(8) Yanez Orozco, I. S.; Mindlin, F. A.; Ma, J.; Wang, B.; Levesque, B.; Spencer, M.; Rezaei Adariani, S.; Hamilton, G.; Ding, F.; Bowen, M. E.; Sanabria, H. Identifying Weak Interdomain Interactions That Stabilize the Supertertiary Structure of the N-Terminal Tandem PDZ Domains of PSD-95. *Nat. Commun.* **2018**, *9* (1), 3724.

(9) Marsh, J. A.; Teichmann, S. A. Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annu. Rev. Biochem.* **2015**, *84* (1), 551−575.

(10) Gershenson, A.; Gosavi, S.; Faccioli, P.; Wintrode, P. L. Successes and Challenges in Simulating the Folding of Large Proteins. *J. Biol. Chem.* **2020**, *295* (1), 15.

(11) Sfriso, P.; Duran-Frigola, M.; Mosca, R.; Emperador, A.; Aloy, P.; Orozco, M. Residues Coevolution Guides the Systematic Identification of Alternative Functional Conformations in Proteins. *Structure* **2016**, *24* (1), 116−126.

(12) Smyth, M. S. X Ray Crystallography. *Molecular Pathology* **2000**, *53* (1), 8.

(13) Marion, D. An Introduction to Biological NMR Spectroscopy. *Molecular & Cellular Proteomics* **2013**, *12* (11), 3006.

(14) Hellenkamp, B.; Schmid, S.; Doroshenko, O.; Opanasyuk, O.; Kühnemuth, R.; Rezaei Adariani, S.; Ambrose, B.; Aznauryan, M.; Barth, A.; et al. Precision and Accuracy of Single-Molecule FRET Measurements—a Multi-Laboratory Benchmark Study. *Nat. Methods* **2018**, *15* (9), 669−676.

(15) Lerner, E.; Cordes, T.; Ingargiola, A.; Alhadid, Y.; Chung, S.; Michalet, X.; Weiss, S. Toward Dynamic Structural Biology: Two Decades of Single-Molecule Förster Resonance Energy Transfer. *Science* **2018**, *359* (6373), eaan1133.

(16) Lerner, E.; Barth, A.; Hendrix, J.; Ambrose, B.; Birkedal, V.; Blanchard, S. C.; Börner, R.; Sung Chung, H.; Cordes, T.; Craggs, T. D. FRET-Based Dynamic Structural Biology: Challenges, Perspectives and an Appeal for Open-Science Practices. *eLife* **2021**, *10*, e60416.

(17) Kalinin, S.; Peulen, T.; Sindbert, S.; Rothwell, P. J.; Berger, S.; Restle, T.; Goody, R. S.; Gohlke, H.; Seidel, C. A. M. A Toolkit and Benchmark Study for FRET-Restrained High-Precision Structural Modeling. *Nat. Methods* **2012**, *9* (12), 1218−1225.

(18) Sasmal, D. K.; Pulido, L. E.; Kasal, S.; Huang, J. Single-Molecule Fluorescence Resonance Energy Transfer in Molecular Biology. *Nanoscale* **2016**, *8* (48), 19928.

(19) Alston, J. J.; Soranno, A.; Holehouse, A. S. Integrating Single-Molecule Spectroscopy and Simulations for the Study of Intrinsically Disordered Proteins. *Methods* **2021**, *193*, 116−135.

(20) Steffen, F. D.; Sigel, R. K. O.; Börner, R. FRETraj: Integrating Single-Molecule Spectroscopy with Molecular Dynamics. *Bioinformatics* **2021**, *37* (21), 3953−3955.

(21) Naudi-Fabra, S.; Blackledge, M.; Milles, S. Synergies of Single Molecule Fluorescence and NMR for the Study of Intrinsically Disordered Proteins. *Biomolecules* **2022**, *12* (1), 27.

(22) Ward, A. B.; Sali, A.; Wilson, I. A. Integrative Structural Biology. *Science (1979)* **2013**, *339* (6122), 913−915.

(23) Naudi-Fabra, S.; Tengo, M.; Jensen, M. R.; Blackledge, M.; Milles, S. Quantitative Description of Intrinsically Disordered Proteins Using Single-Molecule FRET, NMR, and SAXS. *J. Am. Chem. Soc.* **2021**, *143* (48), 20109−20121.

(24) Fuertes, G.; Banterle, N.; Ruff, K. M.; Chowdhury, A.; Mercadante, D.; Koehler, C.; Kachala, M.; Estrada Girona, G.; Milles, S.; Mishra, A.; Onck, P. R.; Gräter, F. Decoupling of Size and Shape Fluctuations in Heteropolymeric Sequences Reconciles Discrepancies in SAXS vs. FRET Measurements. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114* (31), E6342−E6351.

(25) Saikia, N.; Yanez-Orozco, I. S.; Qiu, R.; Hao, P.; Milikisiyants, S.; Ou, E.; Hamilton, G. L.; Weninger, K. R.; Smirnova, T. I.; Sanabria, H.; Ding, F. Integrative Structural Dynamics Probing of the Conformational Heterogeneity in Synaptosomal-Associated Protein 25. *Cell Rep. Phys. Sci.* **2021**, *2* (11), 100616.

(26) Hamilton, G. L.; Saikia, N.; Basak, S.; Welcome, F. S.; Wu, F.; Kubiak, J.; Zhang, C.; Hao, Y.; Seidel, C. A.; Ding, F. Fuzzy Supertertiary Interactions within PSD-95 Enable Ligand Binding. *eLife* **2022**, *11*, e77242.

(27) Sali, A.; Berman, H. M.; Schwede, T.; Trewhella, J.; Kleywegt, G.; Burley, S. K.; Markley, J.; Nakamura, H.; Adams, P.; Bonvin, A.; et al. Outcome of the First WwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **2015**, *23* (7), 1156−1167.

(28) Barth, A.; Opanasyuk, O.; Peulen, T.-O.; Felekyan, S.; Kalinin, S.; Sanabria, H.; Seidel, C. A. M. Unraveling Multi-State Molecular Dynamics in Single-Molecule FRET Experiments. I. Theory of FRET-Lines. *J. Chem. Phys.* **2022**, *156* (14), 141501.

(29) Dimura, M.; Peulen, T.-O.; Sanabria, H.; Rodnin, D.; Hemmen, K.; Hanke, C. A.; Seidel, C. A. M.; Gohlke, H. Automated and Optimally FRET-Assisted Structural Modeling. *Nat. Commun.* **2020**, *11* (1), 5394.

(30) Ambrose, B.; Baxter, J. M.; Cully, J.; Willmott, M.; Steele, E. M.; Bateman, B. C.; Martin-Fernandez, M. L.; Cadby, A.; Shewring, J.; Aaldering, M.; Craggs, T. D. The SmfBox Is an Open-Source Platform for Single-Molecule FRET. *Nat. Commun.* **2020**, *11* (1), 5641.

(31) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108* (49), E1293−E1301.

(32) Dill, K. A.; MacCallum, J. L. The Protein-Folding Problem, 50 Years On. *Science (1979)* **2012**, *338* (6110), 1042−1046.

(33) Hopf, T. A.; Colwell, L. J.; Sheridan, R.; Rost, B.; Sander, C.; Marks, D. S. Three-Dimensional Structures of Membrane Proteins from Genomic Sequencing. *Cell* **2012**, *149* (7), 1607−1621.

(34) Marks, D. S.; Colwell, L. J.; Sheridan, R.; Hopf, T. A.; Pagnani, A.; Zecchina, R.; Sander, C. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLoS One* **2011**, *6* (12), No. e28766.

(35) Sułkowska, J. I.; Morcos, F.; Weigt, M.; Hwa, T.; Onuchic, J. N. Genomics-Aided Structure Prediction. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (26), 10340−10345.

(36) Kamisetty, H.; Ovchinnikov, S.; Baker, D. Assessing the Utility of Coevolution-Based Residue-Residue Contact Predictions in a Sequence- and Structure-Rich Era. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (39), 15674−15679.

(37) Galaz-Davison, P.; Ferreiro, D. U.; Ramírez-Sarmiento, C. A. Coevolution-derived Native and Non-native Contacts Determine the Emergence of a Novel Fold in a Universally Conserved Family of Transcription Factors. *Protein Sci.* **2022**, *31* (6), e4337.

(38) Sutto, L.; Marsili, S.; Valencia, A.; Gervasio, F. L. From Residue Coevolution to Protein Conformational Ensembles and Functional Dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (44), 13567−13572.

(39) Dago, A. E.; Schug, A.; Procaccini, A.; Hoch, J. A.; Weigt, M.; Szurmant, H. Structural Basis of Histidine Kinase Autophosphorylation Deduced by Integrating Genomics, Molecular Dynamics, and Mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (26), E1733−E1742.

(40) Schug, A.; Weigt, M.; Hoch, J. A.; Onuchic, J. N.; Hwa, T.; Szurmant, H. *Computational Modeling of Phosphotransfer Complexes in Two-Component Signaling* **2010**, *471*, 43−58.

(41) Weigt, M.; White, R. A.; Szurmant, H.; Hoch, J. A.; Hwa, T. Identification of Direct Residue Contacts in Protein-Protein Interaction by Message Passing. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106* (1), 67−72.

(42) Ovchinnikov, S.; Kamisetty, H.; Baker, D. Robust and Accurate Prediction of Residue-Residue Interactions across Protein Interfaces Using Evolutionary Information. *eLife* **2014**, *3*, No. e02030.

(43) Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A. The Pfam Protein Families Database: Towards a More Sustainable Future. *Nucleic Acids Res.* **2016**, *44*, D279.

(44) Finn, R. D.; Clements, J.; Eddy, S. R. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res.* **2011**, *39* (suppl), W29−W37.

(45) Eddy, S. R. Profile Hidden Markov Models. *Bioinformatics* **1998**, *14* (9), 755−763.

(46) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235.

(47) Morcos, F.; Jana, B.; Hwa, T.; Onuchic, J. N. Coevolutionary Signals across Protein Lineages Help Capture Multiple Protein Conformations. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110* (51), 20533.

(48) Lammert, H.; Schug, A.; Onuchic, J. N. Robustness and Generalization of Structure-Based Models for Protein Folding and Function. *Proteins: Struct., Funct., Bioinf.* **2009**, *77* (4), 881.

(49) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput. Biol.* **2016**, *12* (3), e1004794.

(50) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and Energetic Factors: What Determines the Structural Details of the Transition State Ensemble and "En-Route" Intermediates for Protein Folding? An Investigation for Small Globular Proteins. *J. Mol. Biol.* **2000**, *298* (5), 937.

(51) Lindahl, E.; Hess, B.; van der Spoel, D. GROMACS 3.0: A Package for Molecular Simulation and Trajectory Analysis. *J. Mol. Model* **2001**, *7* (8), 306.

(52) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph* **1996**, *14* (1), 33.

(53) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera? A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25* (13), 1605.

(54) Opanasyuk, O.; Barth, A.; Peulen, T.-O.; Felekyan, S.; Kalinin, S.; Sanabria, H.; Seidel, C. A. M. Unraveling Multi-State Molecular Dynamics in Single-Molecule FRET Experiments. II. Quantitative Analysis of Multi-State Kinetic Networks. *J. Chem. Phys.* **2022**, *157* (3), 031501.

(55) Schrödinger, L.; DeLano, W.*PyMOL*; Schrödinger, 2020.

(56) Chowdhury, F. N.; Kolber, Z. S.; Barkley, M. D. Monte Carlo Convolution Method for Simulation and Analysis of Fluorescence Decay Data. *Rev. Sci. Instrum.* **1991**, *62* (1), 47−52.

(57) Dix, J. A.; Hom, E. F. Y.; Verkman, A. S. Fluorescence Correlation Spectroscopy Simulations of Photophysical Phenomena and Molecular Interactions: A Molecular Dynamics/Monte Carlo Approach. *J. Phys. Chem. B* **2006**, *110* (4), 1896−1906.

(58) Fanning, A. S.; Anderson, J. M. Protein-Protein Interactions: PDZ Domain Networks. *Curr. Biol.* **1996**, *6* (11), 1385.

(59) Toto, A.; Pedersen, S. W.; Karlsson, O. A.; Moran, G. E.; Andersson, E.; Chi, C. N.; Strømgaard, K.; Gianni, S.; Jemth, P. Ligand Binding to the PDZ Domains of Postsynaptic Density Protein 95. *Protein Eng., Des. Sel.* **2016**, *29* (5), 169.

(60) Bach, A.; Clausen, B. H.; Møller, M.; Vestergaard, B.; Chi, C. N.; Round, A.; Sorensen, P. L.; Nissen, K. B.; Kastrup, J. S.; Gajhede, M.; et al. A High-Affinity, Dimeric Inhibitor of PSD-95 Bivalently Interacts with PDZ1−2 and Protects against Ischemic Brain Damage. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (9), 3317.

(61) Cook, D. J.; Teves, L.; Tymianski, M. Treatment of Stroke with a PSD-95 Inhibitor in the Gyrencephalic Primate Brain. *Nature* **2012**, *483* (7388), 213.

(62) Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins: Struct., Funct., Bioinf.* **2004**, *57* (4), 702−710.