Received: 12 July 2017 | Revised: 26 December 2017 | Accepted: 23 January 2018

DOI: 10.1002/prot.25468

RESEARCH ARTICLE

WILEY PROTEINS

Investigating energy-based pool structure selection in the structure ensemble modeling with experimental distance constraints: The example from a multidomain protein Pub1

Guanhua Zhu[1] | Wei Liu[2] | Chenglong Bao[3,4] | Dudu Tong[1] | Hui Ji[3] | Zuowei Shen[3] | Daiwen Yang[2] | Lanyuan Lu[1]

[1]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore

[2]Department of Biological Sciences, National University of Singapore, 14 Science Drive 4, Singapore 117543, Singapore

[3]Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076, Singapore

[4]Yau Mathematical Sciences Center, Tsinghua University, Haidian District, Beijing 100084, China

Correspondence
YLanyuan Lu, School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore.
Email: LYLU@ntu.edu.sg

Funding information
Ministry of Education, Singapore, Grant/Award Numbers: MOE2014-T2-1–065, MOE2012-T3-1–008

Abstract

The structural variations of multidomain proteins with flexible parts mediate many biological processes, and a structure ensemble can be determined by selecting a weighted combination of representative structures from a simulated structure pool, producing the best fit to experimental constraints such as interatomic distance. In this study, a hybrid structure-based and physics-based atomistic force field with an efficient sampling strategy is adopted to simulate a model di-domain protein against experimental paramagnetic relaxation enhancement (PRE) data that correspond to distance constraints. The molecular dynamics simulations produce a wide range of conformations depicted on a protein energy landscape. Subsequently, a conformational ensemble recovered with low-energy structures and the minimum-size restraint is identified in good agreement with experimental PRE rates, and the result is also supported by chemical shift perturbations and small-angle X-ray scattering data. It is illustrated that the regularizations of energy and ensemble-size prevent an arbitrary interpretation of protein conformations. Moreover, energy is found to serve as a critical control to refine the structure pool and prevent data overfitting, because the absence of energy regularization exposes ensemble construction to the noise from high-energy structures and causes a more ambiguous representation of protein conformations. Finally, we perform structure-ensemble optimizations with a topology-based structure pool, to enhance the understanding on the ensemble results from different sources of pool candidates.

KEYWORDS

energy regularization, molecular dynamics simulation, multidomain protein, paramagnetic relaxation enhancement, structure ensemble modeling

1 | INTRODUCTION

Multidomain proteins and multiprotein complexes with flexible polypeptide segments mediate many biological functions. Nuclear magnetic resonance (NMR) spectroscopy methods provide valuable atomic structural information of biomolecules in solution, and the NMR technique of the paramagnetic relaxation enhancement (PRE)[1,2] experiment is a powerful method for the structural characterization of protein–protein interactions. PRE is able to provide long-range distance information as long as 10∼35 Å, thanks to strong PRE effects contributed by the large magnetic moment of unpaired electron.[3] This unique feature permits PRE experiment as a powerful method for structural characterization of multiprotein complexes[4–8] and DNA–protein complexes[9–11] in solution. Amounts of articles have been published on the structural determination of biomolecules using PRE, whereas a common limitation is the unavailability of a single structural representation satisfying all experimental PRE data. The distance information measured by PRE in many cases involves more than one well-structured unit (domain or protein), and these average distances typically reflect the structure ensemble of various binding states of the units. In some cases, a single structure satisfying experimental PRE data might be identified by appropriate approaches such as simulated annealing[12]; however, this structure may not reflect a real physical structure nor the real conformational space.[3] Therefore, the structural characterization in such cases has largely

depended on ensemble representations constituted by physically reasonable conformations.[3,13]

Restrained molecular simulation is a widely used tool to interpret distance information from experiments.[14] However, although it is possible to directly drive the simulated protein to a state consistent with the interatomic distances,[15–23] similar methods are usually not appropriate for the systems of highly mobile protein–protein interactions, considering the large distance fluctuations and multiple existing binding states. For systems with multiple conformations, most computational investigations focused on generating a pool of protein conformations, which can be subsequently adopted for the structure ensemble optimization with experimental data. However, the determination of conformational ensemble using experimental distance constraint like PRE or other solution experimental data such as small-angle X-ray scattering (SAXS) and residual dipolar couplings (RDCs) remains a challenge for flexible biomolecules, because there might exist multiple solutions achieving an equal fit to experimental observables, known as the ill-posed nature of the problem. Here, the general treatment spirit is imposing extra regularizations to drive the solution to be of desired properties. A common strategy to reach a reliable solution is limiting the total number of conformations within the ensemble to avoid an ambiguous interpretation of experimental data.[3] The approaches based on the ensemble determination with a small number of structures include Minimal Ensemble Search (MES)[24] using a genetic algorithm, Sample and Select (SAS)[25] using a simulated annealing procedure, and other works.[26–28] Besides, the agreement between experimental observables and calculated quantities was also investigated as a function of ensemble size for a di-domain protein U2AF65,[29] where the ensemble construction approach ASTEROIDS[30,31] employed a sampling method with a statistical coil model[32] and a genetic algorithm for optimization. The intuitive idea of using a minimal number of representative structures in PRE fitting can be formulated as the so called L0-norm based mathematic regularization, and a state-of-the-art optimization algorithm Sparse Ensemble Selection was implemented to tackle the problem,[33] linking the structure-fitting issue to the frontier of applied mathematics. In some studies, the populations of the selected conformers were modified to match the experimental observables using another type of regularization methods based on maximum entropy.[28,34,35] Recently, structure energy was proposed as a physics-based regularization term to preclude spurious conformation representations and has been successfully applied to a multidomain protein of dengue virus nonstructure protein 5.[36]

A straightforward computational approach to study protein–protein or domain–domain interactions is brute-force simulation with an all-atom force field.[37,38] One bottleneck for brute-force atomistic simulation is the inadequate sampling caused by the restriction of a short timescale and the inherent weaknesses in force-field parameterizations.[39–42] While advanced simulation methods can be implemented to speed up the dynamics,[43] it is usually very difficult to generate an equilibrated conformational ensemble in an atomistic molecular dynamics (MD) simulation, which explores all possible binding states for protein–protein interactions. Besides, an individual protein or a folded domain is typically believed to preserve its conformational stability in protein–protein interactions,[44,45] while the entire complex could adopt large conformational changes. Therefore, individual proteins or domains are well suited to be represented by using rigid bodies or Gō-like models[46,47] that utilize a structure-based potential adopting the native structure as the global energy minimum. Indeed, recent molecular dynamics simulations[48–50] with coarse-grained (CG) representations of proteins adopted the simulation strategy to separate inter and intra protein interactions. In particular, Kim et al.[27] developed a coarse-grained model for the simulation of multi-protein complexes to fit experimental PRE rates. An alternative method, protein–protein docking, considers individual proteins or folded domains as rigid bodies and searches possible conformational arrangements of the complex. A number of PRE-based algorithms have been devised to extract a structure ensemble from docking results.[7,51,52] This category of computational methods may also include any conformational searching algorithms based on rigid-body representations of single protein units. The main difference between the docking and molecular dynamics methods is that the latter employs a more sophisticated potential energy function that incorporates the full or partial flexibility of proteins. For multidomain proteins and protein complexes, flexible loops in protein(s) and flexible linkers connecting domains usually serve an important role in the protein–protein interaction interface, but they are unstructured in solution due to their high flexibility.[53,54] Although some improved protein–protein docking approaches have incorporated the flexibility of side chains[55] and even backbones,[56] in most cases, MD simulations with a conventional force field produce a structure ensemble of multidomain proteins with flexible segments that might be more sophisticated in physical interactions.

In this work, we investigate the structural characterization of a multi-domain protein by using the published experimental PRE rates[57] and molecular dynamics simulations at atomistic resolution. The target protein is the first two RNA recognition motif domains of poly(U)-binding protein (PubRRM12) in *Saccharomyces cerevisiae*. Poly(U)-binding protein (Pub1) is a major nuclear and cytoplasmic protein that has been suggested as a regulator of cellular mRNA decay.[58] Here, PubRRM12 is of interest as a model system to study multi-domain proteins owing to its structural feature of the two folded domains connected by a flexible linker. The NOE experiments have determined the conformation of individual domains, while PRE experiments give a set of PRE rates corresponding to long-range distances between two domains indicating the large conformational flexibility of the entire complex. Inspired by previous efforts to combine Gō-like model and atomistic force field,[59,60] we develop a hybrid structure-based atomistic model for simulating multidomain proteins. The model adopts an atomistic structure-based potential to model the stable intradomain conformation and meanwhile takes the advantage of accurate empirical all-atom force field to study domain–domain interactions. Restrained molecular dynamics simulations with a sampling strategy are performed to enhance sampling. The resulting simulation structures are clustered and reweighted to fit the experimental PRE rates. Our PRE-based simulation study differs from previous investigations in a few aspects: First, to the best of our knowledge, it is the first time that the folding energy landscape of a di-domain protein is depicted using an atomistic force

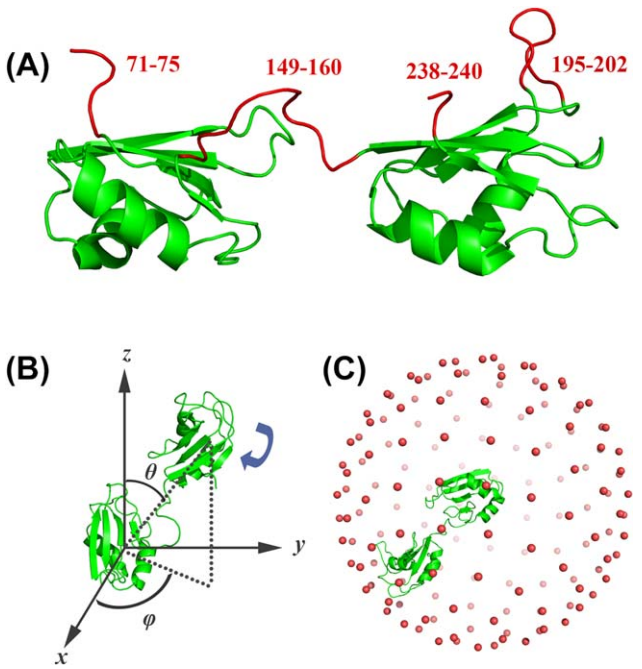

FIGURE 1 Structure representation of PubRRM12 and its domain-domain orientation. A, Rigid (green) and flexible (red) parts of PubRRM12. Amino acids in flexible structures are indicated in red numbers. B, Domain-domain orientation described by angles φ and θ. Self-rotation (blue arrow) of the individual domain is not included in the spherical coordinates. C, 160 vectors uniformly dividing the entire space. RRM1 is placed in the center and RRM2 is pulled away toward the position (red points) each vector points at [Color figure can be viewed at wileyonlinelibrary.com]

field and validated by PRE data. Second, we use the simulated energetic data to exclude unphysical conformations above certain high energy threshold, with a complementary use of small-ensemble regularization that is commonly implemented in PRE studies. We systematically investigate the regularization effect of structure energy and ensemble size in ensemble construction. Finally, we examine the ensemble construction with a topology-based pool, which can enhance our understanding on the structure ensemble results from different sources of pool structures.

2 | METHODS

2.1 | Protein system

The studied protein is the first two RNA recognition motif (RRM) domains of Pub1 (PubRRM12). The initial structures of individual domains were from NMR experiments[57]. The NMR structures of individual domains are basically same as the crystal structures,[58] with pairwise backbone RMSD smaller than 1 Å. The amino acids M107, H123, N148, S190, and N218 in the wild-type protein were mutated to Cysteine. Each cysteine was attached with a nitroxide spin label MTSL (S-(2, 2,5,5-tetramethyl-2,5-dihydro-1H-pyrrol-3-yl)methyl methanesulfonothioate). PRE experiments give the magnitude of the PRE rate proportional to $\langle r^{-6}\rangle$, where r is the distance between backbone amide hydrogen and the oxygen in the attached MTSL. The single-domain structures of two RRM domains are determined by NMR experiments.[57] The partial charges over the MTSL atoms are calculated by Gaussian 09.[61]

2.2 | Potential energy function

In this work, we developed a structure-based atomistic model combined with empirical force field to study the multidomain protein. In our case, the model on atomistic resolution may be more suitable than coarse-grained ones considering the high sensitivity of PRE magnitude proportional to $\langle r^{-6}\rangle$. The overall energy function includes bonded and nonbonded interactions:

$$V = V_{bonded} + V_{nonbonded} \tag{1}$$

We first sort the structure of the multi-domain complex into rigid parts and flexible parts. Usually, folded single domains and proteins can maintain their stable conformations in solution, considered as rigid parts. The rigid parts are subjected to Gō-type potentials with their native structures as the global potential energy minima. The flexible parts refer to the flexible loops and linkers without stable conformations identified in experiments. This treatment explicitly distinguishing rigid parts from flexible parts discriminate our approach from rigid body-based methods. For PubRRM12, experiments[57] identified that the individual domains can be treated as rigid parts except one flexible loop 195–202 and two terminal loops 71–75 and 238–240 (Figure 1A). Besides, the linker from residue 149 to 160 is a flexible part. Finally, the cysteine residues labeled with MTSL are also treated as flexible structures as the local structures are not determined in NMR experiments.

The energy function for bonded interactions is shown in Equation 2:

$$\begin{aligned} V_{bonded} = & \sum_{bond} k_b (b - b_0)^2 \\ & + \sum_{angle} k_\theta (\theta - \theta_0)^2 \\ & + \sum_{\substack{improper/\\ planar}} k_\chi (\chi - \chi_0)^2 \\ & + \sum_{dihedral} k_\varphi \left\{ [1 - \cos(\varphi - \varphi_0)] + \frac{1}{2}[1 - \cos 3(\varphi - \varphi_0)] \right\} \end{aligned} \tag{2}$$

The bonded interactions include the energy terms of bonds, angles, improper and normal dihedrals with strength k_b, k_θ, k_χ, and k_φ, respectively. The term of improper dihedral angles maintains backbone chirality and side-chain planarity. b, θ, χ, and φ are the instantaneous bond distances, angles, improper, and normal dihedral angles; b_0, θ_0, χ^0, and φ_0 represent their equilibrium positions. The rigid parts adopt the equilibrium positions derived from the native structure and the uniform interaction strength for each interaction type, specifically, k_b = 2,000 kJ/(mol×Å^2), k_θ = 150 kJ/(mol×rad2), k_χ = 5 kJ/(mol×rad2), and k_φ = 8 kJ/mol. For the flexible parts, the equilibrium positions and energy strengths are taken from the force field Amber99SB.[62] In the hybrid force field, the strengths of the structure-based interactions are

parameterized to be comparable to their Amber force field counterparts.

The nonbonded energy function is formulated as follows:

$$V_{nonbonded} = V_{ff} + \sum_{native} \varepsilon \left[\left(\frac{r_{0ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{0ij}}{r_{ij}} \right)^{6} \right] \quad (3)$$

where V_{ff} stands for the conventional force-field based non-bonded potential including Lennard-Jones term $V_{ff\text{-}LJ}$ and electrostatic term $V_{ff\text{-}ele}$. $V_{ff\text{-}LJ}$ adopts a conventional LJ pair potential form $V_{ff-LJ} = \sum_{ij} \left(A_{ij}/r_{ij}^{12} - C_{ij}/r_{ij}^{6} \right)$, where parameters A_{ij} and C_{ij} for atom type i and j separated by distance r_{ij} are derived from the atomistic force field. For computational efficiency, no explicit water molecules are simulated. The electrostatic energy term $V_{ff\text{-}ele}$ is given by a Debye–Hückel-type potential $V_{ff-ele} = \sum_{ij} q_i q_j \exp\left(-r_{ij}/\xi\right)/4\pi r_{ij} D$ to describe the shielding effect of the environment between two atoms i and j separated by distance r_{ij}. Avoiding simulating explicit solvent molecules substantially enhances the computational efficiency. In our case, the dominant component of inter-domain interactions comes from the direct van der Waals and electrostatic forces between the binding interfaces. Thus, we consider that a simple dielectric constant is sufficient to describe the basic interaction features and fulfill the task of excluding conformations with very high energy values. The use of a proper dielectric constant is controversial depending on the model used.[63] Here, the dielectric constant 4 is chosen to reflect the dielectric environment in domain-domain interaction interface, in agreement with many studies in the literature.[64–69] $\xi = 14$ Å is the Debye screening length mimicking the screening effect at 20 mM NaCl and 20 mM sodium phosphate salt concentration in experiment. Apart from V_{ff}, an additional two-body structure-based potential term is added to maintain the conformational stability of individual domains. This additional term assigns a Lennard-Jones 6–12 potential over native pairs of heavy atom i and j with respect to their equilibrium position r_{0ij} from the native structure. Native pairs, generated by the method of shadow map,[70] refer to the pairs of heavy atoms that are spatially close located in the same domain excluding flexible structures. These native contacts calculated in this additional term would be exempted from V_{ff} term. A survey of different energy strength was performed, leading a choice of $\varepsilon = 50$ kJ/mol.

2.3 | Sampling and simulation details

The outline of the sampling strategy is a combination of a "pull–push" method for sampling domain–domain orientations and high-temperature MD simulations for sampling self-rotation of individual domains. In our work, the first two RRM domains of Pub1 are marked as RRM1 and RRM2, connected by a flexible linker. The sampling strategy is implemented in three steps. Firstly, the orientation between two domains, considered as mass points placed at their centers of mass (COM), is described by spherical coordinates: angles φ and θ (Figure 1B), ignoring the degrees of freedom of the self-rotation of an individual domain. 160 vectors were generated that uniformly divide the entire space by using the method in literature[71] (Figure 1C). RRM1 was frozen as a rigid body, while RRM2 was pulled away from RRM1 in an MD simulation toward the position each vector pointed at. Next, the resultant 160 conformations after the pull step were subjected to 160 independent high-temperature MD simulations at 1000 K to sample the self-rotation of individual domains. RRM1 was kept frozen and the COM position of RRM2 was fixed by imposing a harmonic potential at its COM position. A time step 0.001 ps was used for a 10 ns simulation, and coordinates were saved every 2 ps. For each replica high-temperature trajectory, clustering was performed on RRM2 using the Gromos algorithm[72] with an RMSD cutoff 3 Å, followed by extracting the middle structure of each cluster. The clustering procedure usually produced 2–6 clusters for each replica trajectory, totally constituting a pool of 611 representative structures. Then, RRM2 was pushed back toward RRM1 along the direction between the COMs of the two domains.

After the push step, the resultant 611 structures were subjected to energy minimization using a steepest descent algorithm with a convergence of maximum force <1000 kJ/mol/nm. The minimized structures were then equilibrated in equilibration MD simulations for 200 ps with a time step of 0.001 ps. The 611 structures after equilibration were the initial structures in the production restrained MD simulations. Restrained MD simulations were performed simultaneously on the 611 initial structures with GROMACS 4.5[73] after the pull–push steps. We implemented weak position restraints on the heavy atoms of RRM1 by imposing a harmonic potential (force constant = 0.005 kJ/mol/nm^2) positioned at each heavy atom. In addition, RRM2 can move only around the direction connecting the domain–domain COMs by adding weak harmonic potentials (force constant = 0.005 kJ/mol/nm^2) along the two perpendicular directions. These weak restraints are imposed for achieving sufficient sampling for each domain-domain orientation. Each replica simulation was carried out for 20 ns at 300 K with the force field Amber99SB,[62] and a time step 0.002 ps was used. The total simulation time is about 12 μs, generating an ensemble of 152,750 conformations.

2.4 | Pool generation and ensemble construction

Clustering was performed on structures from the simulated trajectories in two steps. First, the conformational space of domain–domain orientations that are depicted by φ and θ angles (Figure 1B) is uniformly divided into 72 grids sized 30°×30°. All the simulated structures were grouped on the basis of domain–domain orientations. Next, in each orientation group, clustering was performed using the Gromos algorithm[72] with a small backbone RMSD cutoff 1 Å considering the high sensitivity of PRE rates proportional to $\langle r^{-6} \rangle$. In clustering, the number of neighbors within the cut-off is counted for each structure. The structure with the largest number of neighbors is identified as a cluster center and removed from the pool of structures with all its neighbors. This procedure is repeated for remaining structures in the pool iteratively. The two-step clustering produced 4,003 conformational clusters, and then the middle structure of each cluster was selected as the representative conformation.

The resultant 4,003 representative structures were subjected to the energy calculation by using the molecular mechanics Poisson–

Boltzmann/surface area (MMPBSA) approach.[74] The energy of each representative structure, termed E_{MMPBSA}, consists of E_{MM} and $G_{solvation}$. E_{MM} is the potential energy contributed by the nonbonded interactions that are calculated on the basis of the molecular mechanics (MM) force field Amber99sb. The intradomain components of E_{MM} are excluded from energy calculations because the structures of individual domains are modeled using a Gō-like model. The solvation free energy term $G_{solvation}$ includes polar and nonpolar solvation free energy components whose calculations were implemented using MMPBSA.py module[75] with Amber 14 and AmberTools 14. The polar solvation free energy was calculated by using the Poisson–Boltzmann (PB) approach[76,77] and the nonpolar solvation free energy was composed of repulsive cavitation and attractive dispersion terms.[78] The dielectric constants were set to 1 and 80 for the interior and exterior of solute, respectively. The total 4003 representative conformations are ranked on the basis of their energy E_{MMPBSA}, followed by extracting low-energy structure pools. For example, top-ranking 10% of total structures constitute a top 10% low-energy pool. Here, the choice of energy calculation method is based on recent favorable results[79] of MMPBSA compared with protein-protein docking scoring functions. Additionally, the conformational space of PubRRM12 characterized by MMPBSA is supported by the experimental PRE data and chemical shift perturbations, as shown in the result section. Finally, we tolerate the possible inaccuracy in the calculated energy values to some extent, because we mainly use the energy data to exclude obviously unphysical conformations, and the limitation of force field is minimal in the ensemble construction approach that combines both experimental and simulation information.

For each representative structure, the pair distances between backbone amide hydrogens and MTSL oxygens were calculated and converted to the relaxation rate Γ_2^{cal} in order to fit the experimental PRE rates. Equation 4 shows the relationship between average distance $\langle r_i^{-6} \rangle$ and PRE rate Γ_2^{cal}:

$$\Gamma_2^{cal}(i) = K \langle r_i^{-6} \rangle \left[4\tau_c^{app} + \frac{3\tau_c^{app}}{1+(\omega_H \tau_c^{app})^2} \right] \quad (4)$$

The constant K is 1.23×10^{-44} m^6/s^2. The proton Larmor frequency is $\omega_H = 5.03 \times 10^9$ rad/s for 800 MHz spectrometers and the apparent correlation time is $\tau_c^{app} = 6 \times 10^{-9}$ s. The consistency between experimental and calculated Γ_2^{cal} for residue i is given by the PRE Q-factor defined in Equation 5.[13] For the spins with Γ_2^{obs} and Γ_2^{cal} whose values are >80 s^{-1}, their values are considered as 80 in the Q-factor computation. The detailed explanation and discussion on Equations 4 and 5 are available in literature.[57]

$$Q = \left[\frac{\sum_i \{\Gamma_2^{obs}(i) - \Gamma_2^{cal}(i)\}^2}{\sum_i \{\Gamma_2^{obs}(i)^2\}} \right]^{1/2} \quad (5)$$

We combined representative structures from the low-energy pool to minimize the PRE Q-factor. One state-of-the-art optimization method Sparse Ensemble Selection (SES)[33] was proposed following the small-ensemble philosophy. In SES, top K nonnegative solutions are returned in each step instead of the best one when adding a new structure in the constructed ensemble. K is a user-defined parameter and a larger K value conducts a more complete search but takes a more expensive computational cost. Inspired by this work that selects multiple top solutions in each step, we used a greedy pursuit method to construct the conformational ensemble pursuing a minimal ensemble size, termed Multiple Q-factor Pursuit (MQP). The MQP procedure is described as follows. (1) In the initial step, all the combinations N $(N-1)/2$ between two structures are taken into consideration where N is the total number of candidate structures, followed by selecting top K solutions of the lowest Q-factor for next step. The total complete search in the first step guarantees a good starting point of optimization. (2) Then, for each of the top K ensembles sized two (consisting of two structures) returned from the previous step, a new structure from the remaining candidate pool of $(N-2)$ structures is added to minimize the Q-factor. There are $K \times (N-2)$ combinations of new-step ensembles sized three, among which the top K solutions of the lowest Q-factors are selected for future steps. The weights of ensemble structures are determined by performing non-negativity least squares (NNLS)[80] with the constraint that the summation of all conformation populations is 100%. (3) The top K four-structure ensembles are determined with the same method as described in the previous step. (4) This procedure evolves to some desired steps.

3 | RESULTS AND DISCUSSION

3.1 | Model validation and sampling efficiency

One purpose of our hybrid potential energy function is to maintain the conformational stability of individual domains by using the structure-based potential. The individual domains of PubRRM12 are considered of high conformational stability in solution. The backbone RMSD calculations of RRM1 and RRM2 against the experimental domain structures were implemented on the ensemble obtained from production simulations. Almost all structures throughout the simulation have a RMSD around 1 Å for each individual domain (Supporting Information, Figure S1), indicating that the rigid feature of each domain can be preserved in this model. In comparison, a single long MD simulation was implemented with the standard Amber force field. Without the structure-based potential, large deformations would happen on individual domains (Supporting Information, Figure S1).

It is usually difficult to generate adequate sampling of a large multi-domain protein system from MD simulations with a conventional all-atom force field. To achieve a more complete search, one advanced sampling technique is the use of a large number of trajectories corresponding to different initial structures.[49,81] In a similar spirit, a pull-push sampling strategy combined with high-temperature MD simulations is employed here to generate multiple initial structures. Production simulations with weak restraints assist the sufficient sampling for each domain–domain orientation. If no restraint is imposed, some replicas might deviate from their domain–domain orientations due to the energetically unfavorable initial conformations, consequently causing the loss of orientation sampling. Furthermore, the weak strengths of restraints permit the flexibility of domain–domain interactions to cover

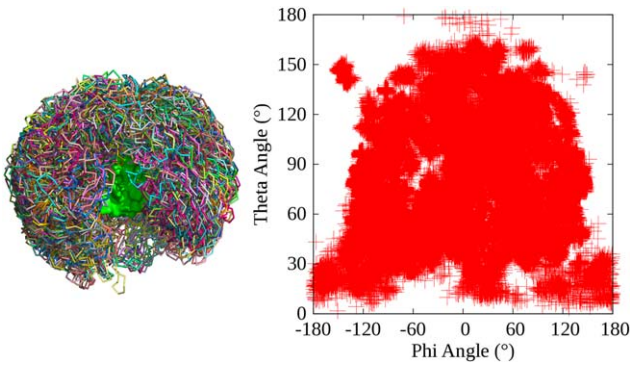

FIGURE 2 Initial structures produced after pull-push step and conformational space sampled by MD simulations starting from these structures. 611 different initial conformations (RRM1 superposition is shown as green surface) are generated (left), from which MD simulations produced a large sampling search defined by domain–domain orientations (right) [Color figure can be viewed at wileyonlinelibrary.com]

adjacent conformational space. It is found that our sampling strategy is able to conduct a more complete search in domain–domain orientations (Figure 2) compared with plain MD simulations and replica-exchange molecular dynamics (REMD) simulations[82] (Supporting Information, Figure S2). The simulations using our sampling methods cover a wide conformational space despite that some orientations are not achievable caused by the physical constraint of the 10-residue linker. In contrast, the simulations without this sampling strategy are highly likely trapped in local minima near the initial conformations, thus their sampling space shows a limited distribution around the starting points. In addition, another way to reveal the conformational space is through the backbone RMSD of the entire complex and domain–domain COM distance. The RMSD versus domain-distance plot (Supporting Information, Figure S3) also suggests a significantly broader conformational space sampled by our methods.

Although our sampling scheme can explore most possible conformations of the two-domain protein, as a tradeoff, the population distribution in the swarm of trajectories does not follow a canonical ensemble distribution. To estimate the relative populations of the conformation clusters, energetic calculations and reweighting optimizations were performed.

3.2 | Energy landscape of interdomain interactions

An energy landscape of domain–domain interactions is a useful tool to reveal the conformational distribution of a multidomain protein in an energy perspective.[49] The two-dimensional energy landscape based on the data describes the domain–domain interactions of PubRRM12, where the depth of a minimum and the number of clusters in the minimum indicate the energetic stabilization and the conformational entropy respectively.[83] In this work, domain–domain orientation defined by angles φ and θ in Figure 1B is selected to present the conformational space of energy landscape. The φ versus θ space depicting domain–domain orientations was divided uniformly into the grids sized 18° × 9°. To organize the great amount of simulated structures, a two-step clustering procedure was performed (Section 2), consequently leading to a total of 4,003 representative conformations. Each representative structure fell into its corresponding grid on the conformational map according to its φ and θ values. In each grid, the structure of the lowest energy E_{MMPBSA} was extracted to represent the energy of that grid. To explicitly display the energy contrast among most areas on the landscape, we performed a search on resetting the zero energy point, consequently leading to the new zero point that is shifted from the grid energy value ranked at the lowest ~46% among all nonempty grids on the landscape.

The energy landscape (Fig. 3) is found as a good illustration to suggest the distribution of energetically favorable domain–domain orientations. Generally, a number of areas of energy minima have been identified on the landscape and each minimum contains multiple low-energy structures. The most striking feature is that these low-energy areas are contiguous on the landscape, constituting a wide energy minimum whose energy level is dramatically lower than other conformational areas. Typically, a relatively smooth "funnel-like" energy landscape for multiprotein complexes is considered as a reflection of guiding individual components to specific binding.[48,84] In our case, the observation of multiple low-energy areas reflects the rugged energy landscape of our multidomain protein, suggesting the absence of strong binding between the two domains.

The chemical shift perturbations from NMR experiments demonstrated the existence of weak interactions between the two domains.[57] Besides, several experimental works on nucleolin[85,86] and sex lethal[87] suggested that RRM domains adopt high conformational flexibility without specific binding in solution and RNA-binding mediates domain–domain interactions. Our energy landscape is further validated by experimental observables in the sections below. Note that some

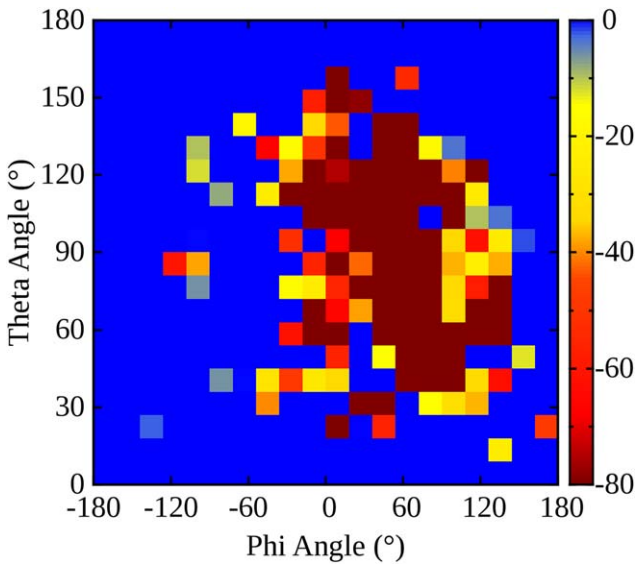

FIGURE 3 Energy landscape of interdomain interactions of PubRRM12. The energy of each domain–domain orientation represented in an 18°×9° gird is indicated in colors defined by the color bar (kcal/mol) [Color figure can be viewed at wileyonlinelibrary.com]

structures located in low-energy areas of domain–domain orientations might still be of high energy because of the self-rotation of individual domains.

3.3 | Ensemble construction with experimental data

The production simulations achieve a better sampling efficiency at the expense of abandoning the rigorous canonical distribution, thus the populations of conformations require a reweighting procedure to fit the experimental PRE rates. Nevertheless, it is not a drawback of our sampling approach as the direct computation of PRE rates without reweighting from an equilibrated MD trajectory is typically formidable in terms of computational expense and force field accuracy. Here, we aim to explore the conformational arrangement of the two domains; therefore, only interdomain PRE rates, referring to the MTSL and the amide hydrogen located in different domains, are used for data fitting. The simulated PRE rates are calculated from pair distances between backbone amide hydrogens and MTSL oxygens (Equation 4). The PRE fitting against experimental data is performed by using a set of structure pools based on energy. Specifically, we apply our MQP method to construct ensembles from top 5%, 10%, 20%, 30%, 40%, 50%, and 100% structure pools, to fit the experimental PRE rates. In the rest of this article, a constructed ensemble optimized from the pool of top a% low-energy structures is called a top a% ensemble.

The solution sparsity of the recovered ensemble is determined by the nature of the biological system. A smooth energy landscape with one or several significant energy minima reflects the specific-binding conformations of multiprotein complexes, while a rugged landscape suggests the absence of specific binding. Without the prior knowledge about the "real" solution such as predetermined conformations from experiments, a small-sized ensemble, empirically no $>$20 structures,[3] was recommended, because larger-sized solutions would cause the ambiguous overinterpretation of experimental data. Besides, another popular approach Ensemble Optimization Method[88] performs ensemble fitting of the SAXS scattering profile from flexible biomolecules and it predefines 20 conformations by default to represent flexible systems. In our study, we adopted MQP approach to construct ensemble sized up to 20 structures, with different energy-level pools.

Figure 4 shows the Q-factor evolution as a function of ensemble size. Q-factors are minimized monotonically for all energy-based ensembles, and the low-energy ensembles that were constructed from low-energy structure pools exhibit a plateau-like pattern within 20 structures. Specifically, the top 5% ensemble produces a converged Q-factor of 0.47 at 11 structures whose fitting quality fails in further improvement with more structures introduced. The enlarged pool of top 10% minimizes the Q-factor to a plateau value of 0.44 using 10 structures. For top 20% and 30% low-energy ensembles, 14 structures are able to lower the Q-factors to 0.39 and 0.35, respectively, which are equal Q-factors as their respective 20-structure counterparts. However, this plateau-like pattern disappears from the top 40% ensemble onward. As shown in Figure 4, the Q-factors of top 40% and 100% ensembles display a continuously decreasing trend within ensemble size up to 20, which imposes a problem of arbitrary ensemble selection

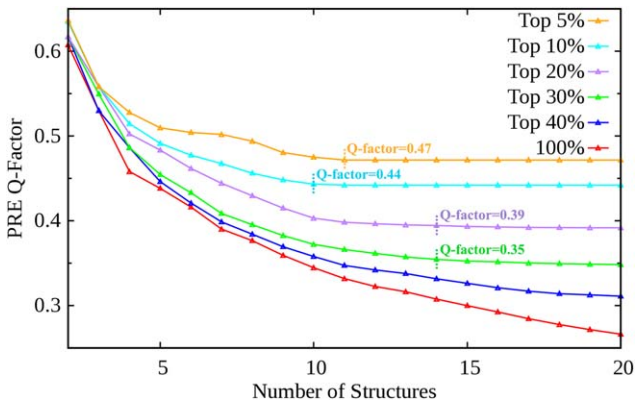

FIGURE 4 PRE Q-factor evolution with ensemble size for top 5%, 10%, 20%, 30%, 40%, and 100% ensembles

without a plateau Q-factor. A looser energy control is likely to expose the constructed ensemble to the noise from high-energy structures, consequently failing to construct a structure ensemble with the small-size restraint. Ensemble-size regularization discourages larger ensembles consisting of $>$20 structures as it would over-interpret the ambiguity of PRE data. Besides ensemble size, energy is proposed as another control factor to prevent overfitting, since energetically unfavorable conformations are of little probability to exist even though they might improve the quality of fit. High-energy structures cause the constructed ensembles doubtable in physics. Therefore, the cross-regularization of ensemble size and energy suggests the use of 14-structure top 30% ensemble to represent the conformation ensemble because of its good performance in both fitting quality and the low-energy control.

Here, the role of energy exclusion in ensemble construction from a given structure pool was emphasized. A structures pool from any computational methods is likely to encounter with significant energy contrast between different sampled poses, while the high-energy one might be selected to represent the conformational space without energy regularization. For example, MD[49,89,90] and Monte Carlo simulations,[27,90] sometimes with replica exchange that is a popular way to enhanced sampling, were adopted to generate candidate structures for a subsequent comparison with experiments. A 200-structure pool from a 40 ns REMD simulation of PubRRM12 was examined using the same method of energetic analyses above, and large energy differences were observed between structures (Supporting Information, Table S1), especially considering its sampled conformations were much incomprehensive (Supporting Information, Figures S2C and S3C). Therefore, energy aspect should be investigated in the ensemble construction from a simulated pool to prevent data overfitting with bad-energy structures.

3.4 | Conformation analysis on constructed ensembles

We investigated the distribution patterns of domain–domain orientation for the constructed low-energy ensembles. The conformational space of domain–domain orientation is depicted in an energy perspective, where a wide major minimum was identified merged by multiple low-energy areas (Figure 3). Figure 5A compares the orientation

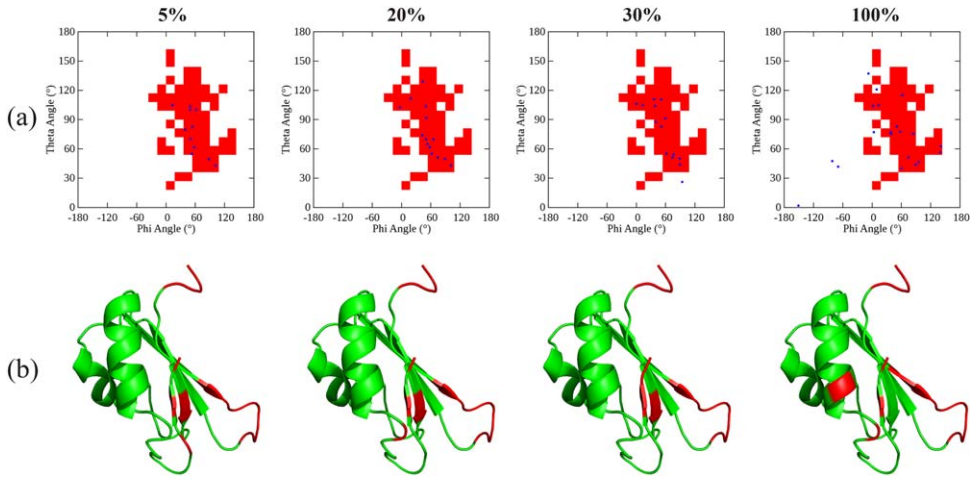

FIGURE 5 Conserved domain–domain orientation and domain contacting interface identified by the constructed low-energy ensembles. A, Domain–domain orientation distribution of ensemble component structures of top 5%, 20%, 30%, and 100% ensembles, shown as blue points. The red area indicates low-energy surface on the energy landscape with a cutoff at top 5% low energy. B, Ribbon representation of RRM1 with contacting residues highlighted in red whose contact numbers are ranked within top 20% among all RRM1 residues [Color figure can be viewed at wileyonlinelibrary.com]

distributions of the structures from top 5%, 20%, 30%, and 100% ensembles. It is noted that the low-energy 5%, 20%, and 30% ensembles are the smallest sized ensembles with the plateau Q-factors as shown in Figure 4. The 100% ensemble is the 20-structure one as it lacks the plateau pattern of the Q-factor evolution with ensemble size. Despite of their different energetic levels, Q-factors and ensemble sizes, their component structures show a highly similar preference on the domain–domain orientation that is located in the energy minimum surface. Therefore, this conserved conformational space is supported by both simulated energy landscape and experimental PRE measurements. Interestingly, the 100% ensemble without energy constraint adopts most of its component structures located in or adjacent to the energy minimum, but three structures ($\varphi < -60°$) of very different orientations located in high-energy area in the energy landscape. It suggests that the major conformational space can be roughly identified from a pure mathematic fit of the experimental PRE observables, but the absence of energy regularization would recover experimental data to a nosier conformational representation caused by high-energy conformations. To further investigate the ensemble structures corresponding to the conserved orientation, the interdomain residue contacts are counted when the residue in RRM1 domain forms a pair-wised contact with the RRM2 residue with their Cα distance <10 Å. The interdomain contacting residues that have most contact numbers are spotted for the contacting interface analysis (Figure 5B). In NMR experiments,[57] some residues with significant chemical shift perturbations were identified indicating these residues are involved in the domain–domain interaction interface. In RRM1, all the residues with significant chemical shift perturbations are located in its β-sheet and adjacent loops, indicating this RRM1 β-sheet face dominates the contacting interface. In RRM2, these residues are from its β-sheet face and also other faces (V171, N172, and Q226) including Q226 in the opposite α-helix face. Compared to the experimental suggestions above, our low-energy ensembles also target the RRM1 β-sheet surface as the predominant domain-domain interacting area (Figure 5B). Actually, the RRM1 β-sheet face does correspond to the major minimum area on the domain–domain orientation map (Figure 5A) because the nature of this φ–θ conformational space is the interaction interface of RRM1. However, the 100% ensemble without energy restraint also targets another novel binding face located at the end of one helix segment in RRM1 domain shown in Figure 5B. This new contacting segment in RRM1 is not supported by chemical shift perturbation experiment, indicating the 100% ensemble without energy restraint may not reflect the real conformational space and molecular interaction. Besides RRM1, The contacting interfaces on RRM2 from all the constructed ensembles are more dispersed than RRM1, just as the chemical shift perturbations indicated.

For comparison, the previous work[57] on PubRRM12 ensemble determination with a genetic algorithm approach used six orthogonal parameters to describe domain–domain relative positions assuming that each domain is a pure rigid body. Aiming for using a minimal-size ensemble to recover experimental measurements, four representative conformations were proposed to reproduce PREs of PubRRM12, and this ensemble also had majorly RRM1 β-sheet facing toward RRM2 domain. Apart from the cross-validation from chemical shift perturbations above, the small-angle X-ray scattering experiments[57] also indicated the existence of multiple conformations of PubRRM12 in solution and derived an approximation of a radius of gyration (R_g) of 18.66 ± 0.53 Å. The calculated R_g of the top 30% low-energy ensemble is 17.44 Å, consistent with the experimental approximation. Thus the constructed low-energy ensembles are further supported by SAXS data in terms of the radius of gyration.

3.5 | The effect of regularizations

In this study, ensemble construction was performed with a combined use of energy-based physics regularization and L0-norm based mathematic regularization. The 20-structure ensemble constructed from the

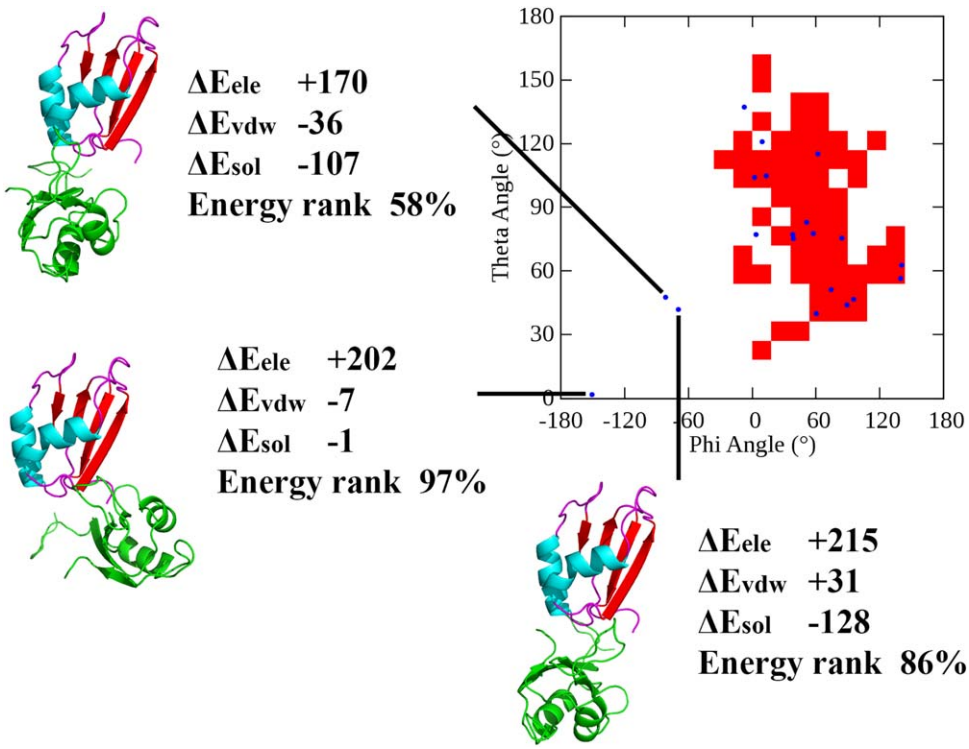

FIGURE 6 Three conformations of minor domain–domain orientations in the constructed 100% ensemble and their energetic profiles. ΔE (kcal/mol) is the difference between the conformation energy and the average energy over all 4003 representative structures. ΔE_{ele}, ΔE_{vdw}, and ΔE_{sol} are the electrostatic, van der Waals, and solvation components of energy difference, respectively. RRM1 structures are colored on the basis of its secondary structure and RRM2 structures are shown in green [Color figure can be viewed at wileyonlinelibrary.com]

100% pool does not take low-energy control in PRE fitting. Although the 100% ensemble can target the major conformational space of interest that is governed by the experimental distance constraints, three component structures of high-energy orientations ($\varphi < -60°$) cause confusion in the conformational representation of PubRRM12. These three structures of novel domain–domain orientations were subjected to energetic analysis to investigate whether there might exist such minor states. A previous work[36] conducted a theoretical investigation and found that a raw numerical fitting without physics-based restraint may cause the wrong conformation representations of the investigated protein. As shown in Figure 6, these three structures turn out to be high-energy structures and their binding modes are highly rejected by electrostatics. Besides, their domain contacting interfaces in RRM1 are novel (Figure 6), totally deviating from RRM1 beta-sheet face that was suggested by chemical shift perturbations. Therefore, the three minor states may be an overfitting result of experimental data without energy regularization because of its violation against both experimental observables and energy landscape.

The restraint type of ensemble size prevents an ambiguous interpretation of experimental data and has been implemented in ensemble recovery in many studies.[24–28,33,57,91] To examine the effect of ensemble size on ensemble construction of PubRRM12, the method of nonnegativity least squares (NNLS)[80] was adopted to reweight candidate structures to fit experimental data best, by assuming the total population to be 100%. With the NNLS optimization scheme, we did not limit the number of structures with nonzero populations, which enables a direct comparison between the most populated conformations suggested by PRE fitting and simulated energetic data. The detailed descriptions of NNLS ensembles of different energy levels are illustrated in Supporting Information, Table S2. The constructed ensembles consist of many dozens of structures after NNLS reweighting scheme, for example top 5% NNLS ensemble of 58 structures, 20% NNLS ensemble of 108 structures, and 30% NNLS ensemble of 121 structures. Despite of different ensemble sizes, these NNLS ensembles have similar numbers of the major conformations (~20) and major conformations constitute a predominant proportion in the constructed ensembles. A major conformation is one with its reweighted population larger than 1% in the constructed NNLS ensemble. However, these NNLS ensembles without size regularization illustrate an arbitrary distribution of domain-domain orientations as shown in Figure 7. This ambiguous interpretation of experimental data causes a spurious representation of PubRRM1 conformations. Interestingly, the major conformations from different NNLS ensembles show a highly similar preference on the domain–domain orientation that is located in the energy minimum (Fig. 7). This is not a surprise because the rough orientation of the two domains is governed by the experimental distance constraints. NNLS fitting without size regularization recovers the major conformational space but introduces too much noise of minor states to interpret the protein conformations.

The "true" structure ensemble of a flexible biomolecule in solution should consist of numerous structures that are mainly distributed in the free energy basins, on the basis of the principles of statistical

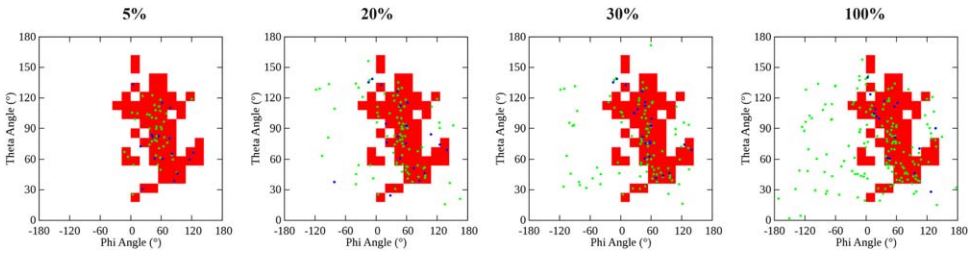

FIGURE 7 Domain–domain orientations of the constructed NNLS ensembles of different energy levels. Domain–domain orientations of ensemble components from reweighted 5%, 20%, 30% low-energy, and 100% NNLS ensembles are shown as blue points for major conformations and green points for minor conformations. The red area indicates low-energy surface on the energy landscape with a cutoff at top 5% low energy [Color figure can be viewed at wileyonlinelibrary.com]

thermodynamics. In practice, a limited amount of structures are employed to represent the biomolecular conformations derived from give experimental data. A good representative ensemble should be able to reflect the true conformational space and molecular interactions. However, considering that experimental data, especially those from low-resolution methods, often contain limited structural information with random and systematic errors, the ensemble recovery from given experimental measurements remains an underdetermined problem for flexible biomolecules. For example, PRE-based calculated ensemble appears artificial due to the limitation of r^{-6} weighted average distances.[92] Data overfitting is highly likely to cause unphysical structures, wrong conformational representatives, and the loss of structural heterogeneity.[93]

The ill-posed problem of ensemble construction of flexible biomolecules might not be fully solved in this work, given that many low-resolution experimental data intrinsically contain insufficient information. However, we show that adding appropriate energy-based constraints can avoid unphysical ensemble solutions. While our computational approach is potentially useful in the interpretation of PRE data, the eventual way to obtain an accurate structural ensemble relies on experimental technical advances for solid high-resolution structural information.

3.6 | Ensemble construction with a topology-based Pool

Pool structure generation and selection are a common strategy in ensemble determination with experimental data, and the candidate pool comes from computational methods based on either molecular interactions[27,34] or molecular topology.[33,52,88] To investigate the potential effect of the pregenerated structure pool on ensemble construction of PubRRM12, we also employed a different source of candidate structures to fit experimental PRE profiles. Unlike the 4003 pool conformations above produced on the basis of atomistic interactions, RANCH[88,94] with a topology-based sampling method for multidomain proteins was adopted to generate a pool of 5,000 structures, followed by adding MTSL molecules to PubRRM12 structures using XPLOR-NIH.[95,96]. The 5000-structure pool size is comparable with our simulated pool. The same MQP optimization scheme above was applied to the topology-based pool. Q-factor evolved with ensemble size up to 20 structures to examine whether the use of a topology-based pool is able to recover experimental observables by a small-sized ensemble. The trend of Q-factor evolution along with structure number has been found as converged at 14 structure of a Q-factor 0.39, equal to its 20-structure counterpart (Figure 8). It is noted that this Q-factor value is larger than the 100% simulated ensemble of 0.27 and even the 30% low-energy simulated ensemble of 0.35 whose candidate pool is much smaller. The reason is that our simulated pool employed molecular simulations to preclude bad-energy structures from pool structures and resulted in a more complete sampling in the important conformational space, such as the energy-favored orientations of PubRRM12. The 14-structure ensemble recovered from the topology-based pool was taken to investigate their conformational space. This topology-based ensemble produces an ensemble average in agreement with experimental PRE measurements, but its most component structures are found being located at the edge of or adjacent to the orientation area of interest (Figure 9). Compared with low-energy simulated ensembles, the topology-based ensemble has a remoter and more dispersed distribution of domain–domain orientations in the experiment-suggested area. Besides, two structures in the topology-based ensemble have different domain contacting interfaces (Figure 9) deviating from the RRM1 beta-sheet, which undermines its interpretation of the conformation space of PubRRM12.

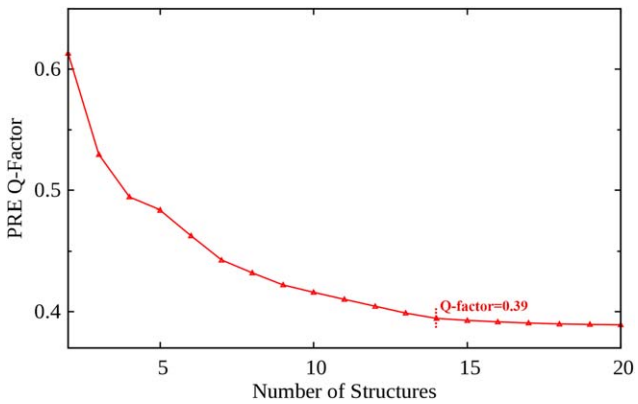

FIGURE 8 PRE Q-factor evolution with ensemble size for the ensemble construction with a topology-based structure pool sized 5000 [Color figure can be viewed at wileyonlinelibrary.com]

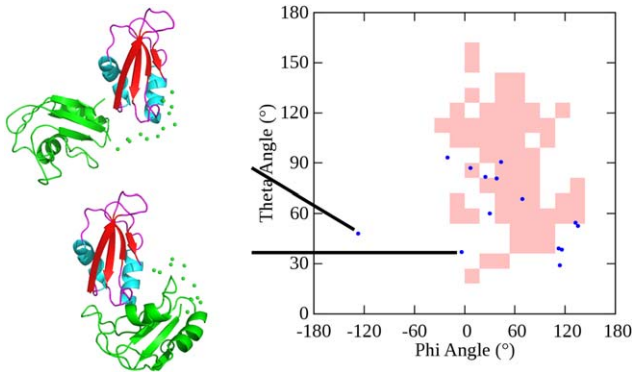

FIGURE 9 Domain–domain orientation distribution (blue points) of 14-structure ensemble recovered from the topology-based pool, two of whose component structures are located away from the low-energy orientation (light red) corresponding to RRM1 beta-sheets [Color figure can be viewed at wileyonlinelibrary.com]

3.7 | The role of linker

Linkers connecting domains usually serve an important role in protein–protein interaction interface,[53,54,97,98] thus we treated them as flexible structures using force field interactions instead of the structure-based potential. One role of linkers is to preclude unphysical conformations in the generated structure pool due to the length restraint of the linker. The linker between RRM1 and RRM2 was broken, followed by MD simulations on PubRRM12 without the physical constraint from the linker. The unphysical conformations (Supporting Information, Figure S4) have been found in the "broken-linker" MD simulations, which should not exit if the linker connects the two domains. The empty space in Supporting Information, Figure S4A depicts the unachievable domain-domain orientations where the two linker-attached residues are in the far-end position. Apart from domain–domain orientation, linker also restricts domain self-rotation (Supporting Information, Figure S4B). Most studies investigate conformational ensemble by generating a pool of protein conformations followed by structure ensemble optimization. The pool generation step should carefully treat the flexible linker considering its significant effect on sampling. To further investigate the linker contribution in domain–domain interactions, the decomposition of interdomain potential energy are examined (Supporting Information, Figure S5). The linker interacts with both domains at the same strength level as the direct interactions between RRM1 and RRM2, implying its important energetic role in domain–domain associations. Linkers are considered as a "scaffold" to precluded unfavorable interdomain interactions.[98] Therefore, a multidomain protein might not be simply treated as two proteins by ignoring the contribution from the conformational arrangement of the linker.

4 | CONCLUSIONS

In this work, we have developed a theoretical pipeline to study multidomain proteins via hybrid structure-based/physics-based atomistic model and pull–push plus high-temperature sampling methods. First, we highlight that our hybrid model is a physically meaningful treatment. It employs a structure-based potential to stabilize the experimentally identified rigid structures and meanwhile ensures the conformational flexibility involving protein–protein interactions. An atomistic force field characterizes interdomain interactions, which permits the more accurate structure analysis with energy calculations using MMPBSA on the atomistic level. Moreover, our MD-based sampling scheme not only precludes bad-energy conformations in contrast with the sampling techniques without molecular interactions, but also conducts a more complete search in the important conformational space. For the critical ensemble-recovery problem, energy is proposed as a control in a complementary use of ensemble-size restraint to regulate the ensemble construction.

One wide energy minimum was detected on the energy landscape of PubRRM12 that shows a high compatibility with experimental PRE observables and chemical shift perturbations, indicating RRM1 β-sheet face dominates the interaction interface. Again, our regularization idea features favorable structures in physics, in a combined use of L0-norm restraint to avoid the overinterpretation of experimental data. The absence of energy constraint would cause a raw numerical fitting and expose the structure ensemble to the high-energy noise, which would cause a more ambiguous conformation representation and undermine the reliability of the constructed ensemble.

Our simulation approach can in principle be applied to any multidomain proteins and multiprotein complexes with dispersed binding states. It would be of particular interest to model a wider variety of experimental measurements, such as data from combined NMR and SAXS studies. Given the recent trend to construct ensemble representations of protein structures by using mixed experimental and simulation methods, our results shed light on the physical interpretation of experimentally derived conformation ensembles.

ACKNOWLEDGMENTS

This research is supported by the Tier 2 grant (MOE2014-T2-1–065) and the Tier 3 grant (MOE2012-T3-1–008) from the Ministry of Education, Singapore. The computational resource from the National Supercomputing Centre of Singapore is acknowledged.

ORCID

Lanyuan Lu http://orcid.org/0000-0003-4808-2431

REFERENCES

[1] Solomon I. Relaxation processes in a system of two spins. *Phys Rev.* 1955;99(2):559

[2] Bloembergen N, Morgan LO. Proton relaxation times in paramagnetic solutions effects of electron spin relaxation. *J Chem Phys.* 1961;34(3):842

[3] Clore GM. Generating accurate contact maps of transient long-range interactions in intrinsically disordered proteins by paramagnetic relaxation enhancement. *Biophys J.* 2013;104(8):1635–1636.

[4] Tang C, Ghirlando R, Clore GM. Visualization of transient ultra-weak protein self-association in solution using paramagnetic relaxation enhancement. *J Am Chem Soc.* 2008;130(12):4048–4056.

[5] Suh JY, Tang C, Clore GM. Role of electrostatic interactions in transient encounter complexes in protein-protein association investigated by paramagnetic relaxation enhancement. *J Am Chem Soc*. 2007;129(43):12954–12955.

[6] Tang C, Schwieters CD, Clore GM. Open-to-closed transition in apo maltose-binding protein observed by paramagnetic NMR. *Nature*. 2007;449(7165):1078–1082.

[7] Volkov AN, Worrall JAR, Holtzmann E, Ubbink M. Solution structure and dynamics of the complex between cytochrome c and cytochrome c peroxidase determined by paramagnetic NMR. *Proc Natl Acad Sci USA*. 2006;103(50):18945–18950.

[8] Tang C, Iwahara J, Clore GM. Visualization of transient encounter complexes in protein-protein association. *Nature*. 2006;444(7117):383–386.

[9] Iwahara J, Clore GM. Detecting transient intermediates in macromolecular binding by paramagnetic NMR. *Nature*. 2006;440(7088):1227–1230.

[10] Iwahara J, Zweckstetter M, Clore GM. NMR structural and kinetic characterization of a homeodomain diffusing and hopping on nonspecific DNA. *Proc Natl Acad Sci USA*. 2006;103(41):15062–15067.

[11] Yadav DK, Lukavsky PJ. NMR solution structure determination of large RNA-protein complexes. *Prog Nucl Magn Reson Spectrosc*. 2016;97:57–81.

[12] Silvestre-Ryan J, Bertoncini CW, Fenwick RB, Esteban-Martin S, Salvatella X. Average conformations determined from PRE data provide high-resolution maps of transient tertiary interactions in disordered proteins. *Biophys J*. 2013;104(8):1740–1751.

[13] Iwahara J, Schwieters CD, Clore GM. Ensemble approach for NMR structure refinement against (1)H paramagnetic relaxation enhancement data arising from a flexible paramagnetic group attached to a macromolecule. *J Am Chem Soc*. 2004;126(18):5879–5896.

[14] Keller B, Christen M, Oostenbrink C, van Gunsteren WF. On using oscillating time-dependent restraints in MD simulation. *J Biomol NMR*. 2007;37(1):1–14.

[15] Allison JR, Rivers RC, Christodoulou JC, Vendruscolo M, Dobson CM. A relationship between the transient structure in the monomeric state and the aggregation propensities of alpha-synuclein and beta-synuclein. *Biochemistry*. 2014;53(46):7170–7183.

[16] Francis CJ, Lindorff-Larsen K, Best RB, Vendruscolo M. Characterization of the residual structure in the unfolded state of the Delta131Delta fragment of staphylococcal nuclease. *Proteins*. 2006;65(1):145–152.

[17] Lindorff-Larsen K, Kristjansdottir S, Teilum K, et al. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *J Am Chem Soc*. 2004;126(10):3291–3299.

[18] Allison JR, Varnai P, Dobson CM, Vendruscolo M. Determination of the free energy landscape of alpha-synuclein using spin label nuclear magnetic resonance measurements. *J Am Chem Soc*. 2009;131(51):18314–18326.

[19] Kristjansdottir S, Lindorff-Larsen K, Fieber W, Dobson CM, Vendruscolo M, Poulsen FM. Formation of native and non-native interactions in ensembles of denatured ACBP molecules from paramagnetic relaxation enhancement studies. *J Mol Biol*. 2005;347(5):1053–1062.

[20] Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM. Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J Am Chem Soc*. 2005;127(2):476–477.

[21] Bertoncini CW, Jung YS, Fernandez CO, et al. Release of long-range tertiary interactions potentiates aggregation of natively unstructured alpha-synuclein. *Proc Natl Acad Sci USA*. 2005;102(5):1430–1435.

[22] Marsh JA, Forman-Kay JD. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *J Mol Biol*. 2009;391(2):359–374.

[23] Gong Z, Gu XH, Guo DC, Wang J, Tang C. Protein structural ensembles visualized by solvent paramagnetic relaxation enhancement. *Angew Chem*. 2017;56(4):1002–1006.

[24] Pelikan M, Hura GL, Hammel M. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen Physiol Biophys*. 2009;28(2):174–189.

[25] Chen Y, Campbell SL, Dokholyan NV. Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophys J*. 2007;93(7):2300–2306.

[26] Huang JR, Grzesiek S. Ensemble calculations of unstructured proteins constrained by RDC and PRE data: a case study of urea-denatured ubiquitin. *J Am Chem Soc*. 2010;132(2):694–705.

[27] Kim YC, Tang C, Clore GM, Hummer G. Replica exchange simulations of transient encounter complexes in protein-protein association. *Proc Natl Acad Sci USA*. 2008;105(35):12855–12860.

[28] Francis DM, Różycki B, Koveal D, Hummer G, Page R, Peti W. Structural basis of p38alpha regulation by hematopoietic tyrosine phosphatase. *Nat Chem Biol*. 2011;7(12):916–924.

[29] Huang JR, Warner LR, Sanchez C, et al. Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. *J Am Chem Soc*. 2014;136(19):7068–7076.

[30] Salmon L, Nodet G, Ozenne V, et al. NMR characterization of long-range order in intrinsically disordered proteins. *J Am Chem Soc*. 2010;132(24):8407–8418.

[31] Guerry P, Salmon L, Mollica L, et al. Mapping the population of protein conformational energy sub-states from NMR dipolar couplings. *Angew Chem Int Ed*. 2013;52(11):3181–3185.

[32] Ozenne V, Bauer F, Salmon L, et al. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*. 2012;28(11):1463–1470.

[33] Berlin K, Castaneda CA, Schneidman-Duhovny D, Sali A, Nava-Tudela A, Fushman D. Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *J Am Chem Soc*. 2013;135(44):16595–16609.

[34] Różycki B, Kim YC, Hummer G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure*. 2011;19(1):109–116.

[35] Choy WY, Forman-Kay JD. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J Mol Biol*. 2001;308(5):1011–1032.

[36] Zhu G, Saw WG, Nalaparaju A, Gruber G, Lu L. Coarse-grained molecular modeling of the solution structure ensemble of dengue virus nonstructural protein 5 with small-angle X-ray scattering intensity. *J Phys Chem B*. 2017;121(10):2252–2264.

[37] Cui Q, Sulea T, Schrag JD, et al. Molecular dynamics-solvated interaction energy studies of protein-protein interactions: the MP1-p14 scaffolding complex. *J Mol Biol*. 2008;379(4):787–802.

[38] Jones MR, Liu C, Wilson AK. Molecular dynamics studies of the protein-protein interactions in inhibitor of kappaB kinase-beta. *J Chem Inf Model*. 2014;54(2):562–572.

[39] Bacci M, Langini C, Vymětal J, Caflisch A, Vitalis A. Focused conformational sampling in proteins. *J Chem Phys*. 2017;147(19):195102

[40] Bernardi RC, Melo MC, Schulten K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim Biophys Acta*. 2015;1850(5):872–877.

[41] Piana S, Klepeis JL, Shaw DE. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Curr Opin Struct Biol*. 2014;24:98–105.

[42] van Gunsteren WF, Daura X, Hansen N, Mark A, Oostenbrink C, Riniker S, Smith L. Validation of molecular simulation: an overview of issues. *Angew Chem*. 2017;

[43] Guerry P, Mollica L, Blackledge M. Mapping protein conformational energy landscapes using NMR and molecular simulation. *ChemPhysChem*. 2013;14(13):3046–3058.

[44] Betts MJ, Sternberg MJ. An analysis of conformational changes on protein-protein association: implications for predictive docking. *Protein Eng*. 1999;12(4):271–283.

[45] Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. *J Mol Biol*. 1999;285(5):2177–2198.

[46] Clementi C, Nymeyer H, Onuchic JN. 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):937–953.

[47] Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY, Onuchic JN. An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins*. 2009;75(2):430–441.

[48] Kim YC, Hummer G. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J Mol Biol*. 2008;375(5):1416–1433.

[49] Ravikumar KM, Huang W, Yang S. Coarse-grained simulations of protein-protein association: an energy landscape perspective. *Biophys J*. 2012;103(4):837–845.

[50] Zhu G, Saw WG, Nalaparaju A, Grüber G, Lu L. Coarse-grained molecular modeling of solution structure ensemble of dengue virus non-structural protein 5 with small-angle X-ray scattering intensity. *J Phys Chem B*. 2017;(in press):DOI: 10.1021/acs.jpcb.1027b00051.

[51] Guan JY, Foerster JM, Drijfhout JW, et al. An ensemble of rapidly interconverting orientations in electrostatic protein-peptide complexes characterized by NMR spectroscopy. *ChemBioChem*. 2014;15(4):556–566.

[52] Kozakov D, Li K, Hall DR, et al. Encounter complexes and dimensionality reduction in protein-protein association. *eLife*. 2014;3:e01370

[53] Ritchie DW. Recent progress and future directions in protein-protein docking. *Curr Protein Pept Sci*. 2008;9(1):1–15.

[54] Swain JF, Dinler G, Sivendran R, Montgomery DL, Stotz M, Gierasch LM. Hsp70 chaperone ligands control domain association via an allosteric mechanism mediated by the interdomain linker. *Mol Cell*. 2007;26(1):27–39.

[55] Vajda S, Camacho CJ. Protein-protein docking: is the glass half-full or half-empty? *Trends Biotechnol*. 2004;22(3):110–116.

[56] Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. *J Mol Biol*. 2007;373(2):503–519.

[57] Liu W, Zhang J, Fan J-S, Tria G, Grüber G, Yang D. A new method for determining structural ensemble of multi-domain protein: application to a RNA binding protein. *Biophys J*. 2016;110(9):1943–1956.

[58] Li H, Shi H, Wang H, et al. Crystal structure of the two N-terminal RRM domains of Pub1 and the poly(U)-binding properties of Pub1. *J Struct Biol*. 2010;171(3):291–297.

[59] Neri M, Anselmi C, Carnevale V, Vargiu AV, Carloni P. Molecular dynamics simulations of outer-membrane protease T from E-coli based on a hybrid coarse-grained/atomistic potential. *J Phys Condens Matter*. 2006;18(14):S347–S355.

[60] Sutto L, Mereu I, Gervasio FL. A hybrid all-atom structure-based model for protein folding and large scale conformational transitions. *J Chem Theory Comput*. 2011;7(12):4208–4217.

[61] Frisch MJ, Trucks GW, Schlegel HB, et al. *Gaussian 09*. Wallingford, CT, USA: Gaussian, Inc.; 2009.

[62] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*. 2006;65(3):712–725.

[63] Schutz CN, Warshel A. What are the dielectric "constants" of proteins and how to validate electrostatic models? *Proteins*. 2001;44(4):400–417.

[64] Yang T, Wu JC, Yan C, et al. Virtual screening using molecular simulations. *Proteins*. 2011;79(6):1940–1951.

[65] Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. *J Comput Chem*. 2004;25(9):1157–1174.

[66] Fogolari F, Zuccato P, Esposito G, Viglino P. Biomolecular electrostatics with the linearized Poisson-Boltzmann equation. *Biophys J*. 1999;76(1 Pt 1):1–16.

[67] Fogolari F, Brigo A, Molinari H. Protocol for MM/PBSA molecular dynamics simulations of proteins. *Biophys J*. 2003;85(1):159–166.

[68] Meyer T, Knapp EW. pKa values in proteins determined by electrostatics applied to molecular dynamics trajectories. *J Chem Theory Comput*. 2015;11(6):2827–2840.

[69] Kumari R, Kumar R, Open Source Drug Discovery C, Lynn A. g_mmpbsa - A GROMACS tool for high-throughput MM-PBSA calculations. *J Chem Inf Model*. 2014;54(7):1951–1962.

[70] Noel JK, Whitford PC, Onuchic JN. The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *J Phys Chem B*. 2012;116(29):8692–8702.

[71] Ponti A. Simulation of magnetic resonance static powder lineshapes: a quantitative assessment of spherical codes. *J Magn Reson*. 1999;138(2):288–297.

[72] Daura X, Gademann K, Jaun B, Seebach D, van Gunsteren WF, Mark AE. Peptide folding: when simulation meets experiment. *Angew Chem Int Ed*. 1999;38(1–2):236–240.

[73] Pronk S, Pall S, Schulz R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*. 2013;29(7):845–854.

[74] Kollman PA, Massova I, Reyes C, et al. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc Chem Res*. 2000;33(12):889–897., 3rd,.

[75] Miller IIIBR, McGee TD, Jr, Swails JM, Homeyer N, Gohlke H, Roitberg AE. MMPBSA. py: an efficient program for end-state free energy calculations. *J Chem Theory Comput*. 2012;8(9):3314–3321.

[76] Honig B, Nicholls A. Classical electrostatics in biology and chemistry. *Science*. 1995;268(5214):1144–1149.

[77] Warwicker J, Watson HC. Calculation of the electric-potential in the active-site cleft due to alpha-helix dipoles. *J Mol Biol*. 1982;157(4):671–679.

[78] Tan C, Tan YH, Luo R. Implicit nonpolar solvent models. *J Phys Chem B*. 2007;111(42):12263–12274.

[79] Chen F, Liu H, Sun H, et al. Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein-

protein binding free energies and re-rank binding poses generated by protein-protein docking. *Phys Chem Chem Phys*. 2016;18(32): 22129–22139.

[80] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge, U.K.: Cambridge University Press; 2007.

[81] Czerminski R, Elber R. Computational studies of ligand diffusion in globins: I. Leghemoglobin. *Proteins*. 1991;10(1):70–80.

[82] Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*. 1999;314(1–2):141–151.

[83] Onuchic JN, Luthey-Schulten Z, Wolynes PG. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem*. 1997;48:545–600.

[84] Levy Y, Cho SS, Onuchic JN, Wolynes PG. A survey of flexible protein binding mechanisms and their transition states using native topology based energy landscapes. *J Mol Biol*. 2005;346(4):1121–1145.

[85] Allain FH, Bouvet P, Dieckmann T, Feigon J. Molecular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. *Embo J*. 2000;19(24):6870–6881.

[86] Allain FH, Gilbert DE, Bouvet P, Feigon J. Solution structure of the two N-terminal RNA-binding domains of nucleolin and NMR study of the interaction with its RNA target. *J Mol Biol*. 2000;303(2):227–241.

[87] Crowder SM, Kanaar R, Rio DC, Alber T. Absence of interdomain contacts in the crystal structure of the RNA recognition motifs of Sex-lethal. *Proc Natl Acad Sci USA*. 1999;96(9):4892–4897.

[88] Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. Structural characterization of flexible proteins using small-angle X-ray scattering. *J Am Chem Soc*. 2007;129(17):5656–5664.

[89] Terakawa T, Higo J, Takada S. Multi-scale ensemble modeling of modular proteins with intrinsically disordered linker regions: application to p53. *Biophys J*. 2014;107(3):721–729.

[90] Metskas LA, Rhoades E. Conformation and dynamics of the troponin I C-terminal domain: combining single-molecule and computational approaches for a disordered protein region. *J Am Chem Soc*. 2015;137(37):11962–11969.

[91] Frank AT, Stelzer AC, Al-Hashimi HM, Andricioaei I. Constructing RNA dynamical ensembles by combining MD and motionally decoupled NMR RDCs: new insights into RNA dynamics and adaptive ligand recognition. *Nucleic Acids Res*. 2009;37(11):3670–3679.

[92] Ganguly D, Chen J. Structural interpretation of paramagnetic relaxation enhancement-derived distances for disordered protein states. *J Mol Biol*. 2009;390(3):467–477.

[93] Allison JR. Using simulation to interpret experimental data in terms of protein conformational ensembles. *Curr Opin Struct Biol*. 2017;43: 79–87.

[94] Petoukhov MV, Svergun DI. Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophys J*. 2005;89(2):1237–1250.

[95] Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM. The Xplor-NIH NMR molecular structure determination package. *J Magn Reson*. 2003;160(1):65–73.

[96] Schwieters CD, Kuszewski JJ, Clore GM. Using Xplor-NIH for NMR molecular structure determination. *Prog Nucl Magn Reson Spectrosc*. 2006;48(1):47–62.

[97] Gokhale RS, Khosla C. Role of linkers in communication between protein modules. *Curr Opin Chem Biol*. 2000;4(1):22–27.

[98] George RA, Heringa J. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng Des Sel*. 2002; 15(11):871–879.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Zhu G, Liu W, Bao C, et al. Investigating energy-based pool structure selection in the structure ensemble modeling with experimental distance constraints: The example from a multidomain protein Pub1. *Proteins*. 2018;00:1–14. https://doi.org/10.1002/prot.25468