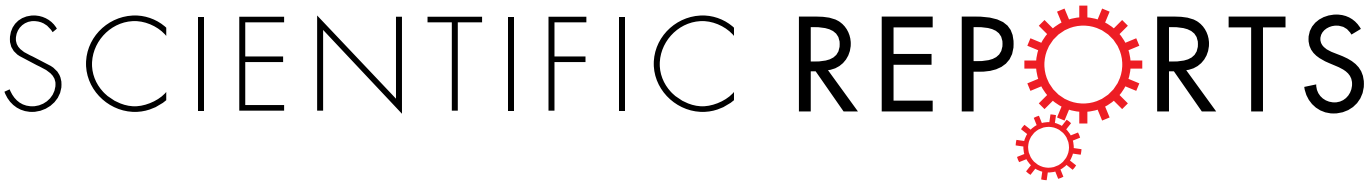

OPEN

SAXS-guided Enhanced Unbiased Sampling for Structure Determination of Proteins and Complexes

Received: 19 March 2018
Accepted: 12 November 2018
Published online: 10 December 2018

Chuankai Zhao[1] & Diwakar Shukla[1,2,3,4]

Molecular simulations can be utilized to predict protein structure ensembles and dynamics, though sufficient sampling of molecular ensembles and identification of key biologically relevant conformations remains challenging. Low-resolution experimental techniques provide valuable structural information on biomolecule at near-native conditions, which are often combined with molecular simulations to determine and refine protein structural ensembles. In this study, we demonstrate how small angle x-ray scattering (SAXS) information can be incorporated in Markov state model-based adaptive sampling strategy to enhance time efficiency of unbiased MD simulations and identify functionally relevant conformations of proteins and complexes. Our results show that using SAXS data combined with additional information, such as thermodynamics and distance restraints, we are able to distinguish otherwise degenerate structures due to the inherent ambiguity of SAXS pattern. We further demonstrate that adaptive sampling guided by SAXS and hybrid information can significantly reduce the computation time required to discover target structures. Overall, our findings demonstrate the potential of this hybrid approach in predicting near-native structures of proteins and complexes. Other low-resolution experimental information can be incorporated in a similar manner to collectively enhance unbiased sampling and improve the accuracy of structure prediction from simulation.

Proteins fold into precise three-dimensional structures to carry out essential cellular functions such as enzyme catalysis[1–5] and signaling[6–12]. To understand protein structure-function relationships, it is crucial to obtain knowledge of not only key biologically relevant functional conformations but also kinetic pathways associated with the conformational change process. Although X-ray crystallography and nuclear magnetic resonance (NMR) techniques can provide high-resolution protein structures, it is often difficult to capture all conformation states of proteins in solution. Complementarily, low-resolution experimental techniques, such as small angle X-ray scattering (SAXS)[13], single molecule fluorescence resonance energy transfer (smFRET)[14] and double electron-electron resonance (DEER)[15] can be utilized to gain insights into the protein conformational states or dynamics in solution. However, due to the low information content of experimental data, these techniques alone are not sufficient to obtain high-resolution protein structures. Instead, additional physical or structural information is required to interpret the information encoded by low-resolution experimental data and prevent overfitting during structure determination and refinement.

Molecular dynamics (MD) simulation is a powerful tool to complement low-resolution experimental data to predict protein structures and ensembles, as well as thermodynamics and kinetics associated with protein function[16,17]. One popular way is to utilize coarse grained MD simulation to generate structural ensemble and then refine the ensemble against experimental data[18–21]. However, coarse grained simulations might not be accurate enough to represent structural ensemble in atomic detail. To improve the predictive accuracy, a recent study utilizes extensive all atom MD simulations to generate a strong prior structural ensemble and incorporates

[1]Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, United States. [2]Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, United States. [3]Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, United States. [4]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL, 61801, United States. Correspondence and requests for materials should be addressed to D.S. (email: diwakar.shukla@shuklagroup.org)

experimental information in a statistical approach to refine the ensemble against experimental data by perturbing the weight of each state in the kinetic model[22]. This method can likely generate the true ensemble under sufficient sampling of molecular ensembles. However, this approach involves *post-hoc* validation of the conformational ensembles based on the extensive simulation data obtained from unbiased MD simulations. It is computationally demanding to generate extensive simulation data for biological processes where conformational dynamics takes place over a millisecond or even longer timescales (i.e. protein folding) or where large biomolecule sizes can greatly hamper the computation speed (i.e. protein-protein association).

To address some of these limitations, low-resolution experimental data, which offers a direct observable to compare simulation with experiment, can be combined with MD sampling process to accelerate discovery of functionally relevant structures of proteins. To achieve this, experimental data is often incorporated as constraints in enhanced sampling algorithms, either through modifying potential functions[23–27] or defining reaction coordinates[28–30], to bias simulations and drive molecules of interests towards conformational states that are consistent with experimental data. Although these methods have improved computational efficiency and succeeded in predicting functional conformations of proteins, they may sacrifice kinetic information for accurate thermodynamics.

In this study, we aim to explore how unbiased MD simulations and low-resolution experimental data could be integrated to obtain accurate conformational ensembles of proteins, while at the same time reducing computational costs. In particular, we explore to quantify the advantage of incorporating low-resolution experimental data in conformational sampling algorithm by evaluating the effectiveness of this approach in terms of both structure prediction accuracy and sampling efficiency. Here, SAXS is used as the source of low-resolution structural information due to its wide spread use in structural characterization of both structured and intrinsically disordered biomolecules in solution, especially for complexes[13,31,32]. SAXS data is presented as a one dimensional scattering curve determined from the spherical averaging of random orientations that a biomolecule can adopt in solution. It remains elusive how the information is distributed over the range of scattering curve, however usually experimental SAXS profiles do not contain more than 10–30 independent points[17,33]. Low-resolution shapes of biomolecules can be reconstructed using SAXS data, and structural features such as radius of gyration (R_g) and maximum diameter (D_{max}) of biomolecules can be extracted by fitting SAXS profile. The low information content leads to the inherent ambiguity of SAXS data that biomolecules with different shape topologies and internal structures can display identical SAXS profiles[31,33,34]. Thanks to the recent advances in SAXS data collection with reduced noises and errors, as well as accurate prediction of SAXS profile from structural models, it becomes possible to harvest the structural information encoded in SAXS data[17,35].

We present an approach that adopts the SAXS information as a seed selection criteria for Markov state model (MSM)-based adaptive sampling[36] in unbiased MD simulations to enhance the sampling of protein dynamics and identify near-native conformational states of proteins. As compared to experimental-guided enhanced sampling algorithms, this approach does not introduce changes to MD potential functions, thereby providing more accurate description of equilibrium protein dynamics. Furthermore, different types of structural information can be incorporated at the same time to collectively enhance the sampling of protein dynamics. In this work, we demonstrate this method in the study of protein folding and protein-protein association. Especially, considering the under-determined nature of SAXS data, we aim to explore what additional information might be needed to combine with SAXS to identify native states of proteins and complexes from MD simulations and to better enhance MD sampling efficiency.

To first demonstrate how SAXS and SAXS-based hybrid information can be used to both predict the native structure of single domain protein and reduce the computation time to discover the target structure, we study the foldings of three proteins, including HP35 double norleucine mutant domain (35 residues)[37], protein G (56 residues)[38] and α3D (73 residues)[38]. Markov state models (MSMs)[39,40] are constructed using the previously published extensive MD simulation datasets. We show that the combination of SAXS and thermodynamic information estimated from the MSMs is sufficient to clearly differentiate structures and recognize the folded structure for the three small single domain proteins. By performing kinetic Monte Carlo sampling[41] using different sampling protocols on the MSMs, we show that SAXS-guided adaptive sampling strategy greatly reduces the simulation cost of reaching the target structure as compared to other sampling methods. We also demonstrate that distance restraints inferred from intramolecular evolutionary couplings (ECs) can be combined with SAXS to further improve the prediction accuracy and sampling efficiency.

Next, we extend the utility of this approach in predicting the protein-protein association pathways based on available structures of individual subunits. We analyze the previously published MD simulations of the association of *E. coli* molybdopterin synthase subunits MoaD and MoaE[42]. Our results suggest that SAXS data must be combined with distance restraints, which are inferred from intermolecular ECs to better distinguish degenerate MoaD-MoaE structures displaying similar SAXS profiles. Using kinetic Monte Carlo sampling on the MSM built using the simulation datasets, we demonstrate that the utility of SAXS data in adaptive sampling can still reduce the computation time to reach the target structure. Furthermore, the sampling efficiency is further enhanced by utilizing both SAXS and distance restraints. Finally, we demonstrate the application of this approach in actual MD simulations to study the association of homodimer of plant hormone receptor PYR1[43]. As in MoaD-MoaE association, by combining SAXS information and distance restraints, we discover a structure that aligns well with the crystal structure (C_α RMSD: 3.18 Å) with a limited sampling time. Our study demonstrates that SAXS-guided adaptive sampling is an efficient approach to predict not only near-native structure ensemble but also transition pathways of conformational changes of proteins from simulation.

Methods

SAXS-guided adaptive sampling. The pipeline of adaptive sampling consists of iteratively running short parallel simulations, clustering the trajectories based on some structural features, and seeding new simulations from certain clusters according to some selection criterion[36]. The key of SAXS-guided adaptive sampling is to incorporate the SAXS information in the selections of seeding structures for iterative sampling. This is achieved by converting the SAXS profile into a SAXS discrepancy scoring function, which measures the degree of similarity between the target experimental or theoretical SAXS profile and the SAXS profile calculated from the structural models of each cluster. By selecting the clusters which are closer to the target, we bias the sampling direction while leaving the energy function unchanged. By iteratively running short parallel simulations, we drive the system of interest towards the target structure while still maintaining accurate thermodynamics and kinetics in the sampling process. The SAXS discrepancy function used in this study is the commonly used reduced χ^2 function (equation 1):

$$\chi^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\mu I_{state}(q_i) - I_{target}(q_i)}{\sigma_{target}(q_i)}\right)^2 \quad (1)$$

where q_i is the momentum transfer ($q = 4\pi\sin\theta/\lambda$, 2θ is the scattering angle and λ is the x-ray wave length), $I_{target}(q_i)$ and $I_{state}(q_i)$ are the scattering intensities of target SAXS profile and each cluster state at q_i, $\sigma_{target}(q_i)$ is the error of target scattering intensity at q_i, N is the total number of data points in the SAXS scattering curves, μ is a scaling factor.

Markov state model. Markov state model is a kinetic network model built from MD simulation datasets to describe the protein conformation space with discrete states and their transition probabilities[39,40]. The discrete states are generated from the clustering of all protein conformations based on some relevant structural metrics. The transition probabilities between these states are estimated by a maximum likelihood analysis of the interstate transition counts from the raw trajectories[44]. For an N state model with a lag time τ, the behavior of any initial probability distribution $\boldsymbol{p}(t)$ over time can be given by (equation 2):

$$\boldsymbol{p}(t + k\tau) = \boldsymbol{p}(t)\boldsymbol{T}(\tau)^k \quad (2)$$

where $\boldsymbol{p}(t)$, $\boldsymbol{p}(t+k\tau)$ are N dimensional vectors of the state probabilities, $\boldsymbol{T}(\tau)$ is the transition probability matrix with each component T_{ij} as the transition probability between state i, j at a period of τ. The eigenvectors of $\boldsymbol{T}(\tau)$ in descending order by eigenvalues represent approximations of the underlying continuous-space propagator, where the first eigenvector is the equilibrium state probability vector $\boldsymbol{w}$[45]. The free energy of each MSM state i (G_i) can be estimated by $G_i = -RTln(w_i)$, where R and T are the gas constant and temperature. Using MSMs, long timescale behavior of protein dynamics can be accurately predicted. All the MSMs in this study were constructed using MSMBuilder 3.4 package[46].

Kinetic Monte Carlo (MC) sampling on MSM. Kinetic MC simulation is a probabilistic method based on MSM transition probability matrix to generate arbitrary trajectory to reveal long term state-to-state protein dynamics[41]. If the initial state is chosen as state i, the probability of a transition from state i to state j over a period of τ is T_{ij}. This is implemented by generating a pseudorandom number between 0 and 1, and taking a cumulative sum of T_{ij} over j ($S_n = \sum_j T_{ij}$). If the random number lies between S_n and S_{n+1}, then the state $n+1$ will be added to the trajectory. All the kinetic MC simulations in this study were conducted using MSMBuilder 3.4 package[46].

Computation of SAXS profiles. For the HP35, the protein G, the α3D domain and the MoaD-MoaE systems, all the SAXS profiles were calculated from the protein structure coordinates using the Crysol software[47] provided in the ATSAS software package[48], version 2.7.2. The SAXS scattering intensities were calculated between 0–0.5 $Å^{-1}$ with 51 points in total. The number of points were chosen larger than the number of Shannon channels, as given by $N_s = q_{max}D_{max}/\pi$, where q_{max} is the maximum scattering vector and D_{max} is the maximum diameter of protein[49,50]. The number of harmonics was set to 40 and the order of Fibonacci grid was set to 18, and the default values of all other parameters were used. In the Crysol, the solvation shell of biomolecule is approximated by a border layer of certain effective thickness with a density differed from the bulk[47]. The contrast of hydration shell was set to 0.03 e/$Å^3$ (default) and the solvent density was set to 0.334 e/$Å^3$ (default).

For the PYR1 homodimer, all the SAXS profiles were calculated from the explicit solvent structural models using the WAXSiS algorithm[51,52]. In the WAXSiS, a spatial envelope that encloses all conformational states of the biomolecule with sufficient distance d, as well as the solvation layer, is defined to allow for the calculation of SAXS profiles from explicit solvent structural ensembles, while at the same time reducing statistical noise and computational cost. The explicit solvent representation effectively considers the structured water pattern within the solute solvation shell, allowing for accurate calculation of the scattering intensities at wide angles[51]. In this study, we calculated the SAXS scattering intensities at q between 0 to 1 $Å^{-1}$ with 101 points (larger than the number of Shannon channels[49,50] $N_s \approx 22$). The envelope distance d was 7 Å, which has been shown to be enough to ensure bulk-like solvent density at envelope surface[51]. The density of the solvent was corrected to match the experimental value 0.334 e/$Å^3$ using the density correction scheme implemented in WAXSiS. For each absolute value of the scattering angle q, 1000 homogeneous scattering vectors q_j ($j = 1, \ldots, 1000$) were used for numerical computation of spherical average scattering intensity $I(q)$. All calculations were performed using the WAXSiS implementation in modified version of the GROMACS simulation software, version 4.62[53].

R_g and D_{max} estimated from the calculated or experimental SAXS profiles were determined using Guinier analysis as implemented in the ATSAS software package[48]. In addition to using the reduced χ^2 function to assess

the similarity between the target and predicted SAXS profiles, we also employed the correlation map method as implemented in the ATSAS software package[48], which does not rely on the estimation of the errors of target SAXS profile. Gaussian random noises were added to the target scattering intensities $I_{target}(q_i)$ to account for the errors while using the correlation map method. For example, Gaussian random noise at q_i was generated by a random number from a normal distribution with mean of 0 and standard deviation of $\sigma_{target}(q_i)$. The randomly generated Gaussian noise was then added to $I_{target}(q_i)$.

Model systems

Folding of single domain proteins. The total MD simulation times for the folding of HP35 double norleucine mutant, protein G, and α3D domain analyzed in this study were approximately 294, 1154, 707 μs respectively[37,38]. The HP35 folding trajectories were clustered using k-centers algorithm based on the root mean square deviations (RMSDs) of all heavy atoms from the HP35 crystal structure (PDB ID: 2F4K[54]), same as in the previous literature[37]. An MSM with 500 states and a lag time τ of 30 ns was constructed. The protein G and α3D domain folding trajectories were clustered using k-means algorithm based on the 100 slowest-relaxing degrees of freedom from linear combinations of all ϕ, ψ and χ_1 dihedral angles using the time-lagged independent component analysis (tICA)[55]. MSMs with 500 states and lag times of 50 ns were constructed. The lag times were chosen based on the convergence of implied timescales (Supplementary Fig. S1a–c). For each state of the MSMs, 100 structures were randomly extracted from the simulation datasets to calculate the SAXS profile of the state.

In order to calculate the SAXS profiles of native states of HP35, protein G and α3D, 50 ns explicit-solvent MD simulations on the experimentally determined structures of HP35, protein G and α3D (PDB IDs: 2F4K[54], 1MI0[56], 2A3D[57]) were performed in Amber14 using the Amber ff14SB force field[58]. The structures were solvated with TIP3P water and Na^+/Cl^- were added to the system with the salt concentration of 0.15 M using AmberTools15. Subsequently, 10000 steps of energy minimization and 2 ns equilibration were performed for each system. Simulations were performed with a 2 fs time step and maintained at 300 K, 1 atm using Berendsen thermo-barostat[59]. The SHAKE algorithm[60] was applied to constrain the length of covalent bonds involving hydrogen atoms. The Particle-mesh Ewald method[61] was used to treat the electrostatic interactions with a 10 Å cutoff distance. 100 snapshots from the 50 ns MD simulations were extracted to calculate the SAXS profiles, and the non-weighted average of the 100 SAXS profiles was calculated to serve as the target SAXS profiles for adaptive sampling. More specifically, at each scattering vector q_i, the average of scattering intensities ($I(q_i)$) and the standard deviation of $I(q_i)$ were determined as $I_{target}(q_i)$ and $\sigma_{target}(q_i)$, respectively. For all states, the SAXS discrepancy values were calculated using the reduced χ^2 function.

Three different sampling strategies: (1) traditional long simulation, (2) random adaptive sampling and (3) SAXS-guided adaptive sampling were employed in kinetic MC sampling to compare their sampling efficiency for all three proteins. The initial state was the expanded unfolded state with the largest RMSD value from target structure. The total sampling times required to reach the predicted folded state from the initial state were calculated for these three sampling schemes. In traditional long simulations, varying number of parallel simulations (10 to 1000) starting from the initial state were run for varying amount of time (1τ to 15τ), and 1500 sets of synthetic simulations were run in total. In random adaptive sampling, 10 parallel trajectories were launched from the initial state and run for varying amount of time (1τ to 15τ). Then 10 new states were randomly chosen from the resulting trajectories as the seeds for next round of sampling. This process iterated for varying number of adaptive rounds (1 to 100) and again, 1500 sets of synthetic simulations were run in total. Lastly, SAXS-guided adaptive sampling was following exactly the same procedures as in random adaptive sampling except 10 seeds in each round were chosen from the states that give the lowest SAXS discrepancy values. The sampling times required to reach the predicted folded state were calculated to quantify the sampling efficiency using different protocols.

Finally, for HP35 and protein G, we also tested the sampling efficiency using SAXS-based hybrid information-guided sampling approach combining both SAXS and distance restraints inferred from intramolecular evolutionarily couplings (ECs). The ECs were extracted using a pseudolikelihood (PLM) method[62] on EVCouplings web server[63] (http://evfold.org) with the default settings. For each MSM state, the distances of top ranked evolutionarily coupled residue pairs (with EC score > 0.3, 10 ECs for HP35, and 7 ECs for protein G, Supplementary Fig. S2) were calculated. The average residue pair distances were used for adaptive sampling. Under SAXS-EC-guided adaptive sampling protocol, we iteratively chose 5 states with the minimal SAXS discrepancy scores and 5 states with the minimal EC residue pair distance for adaptive sampling. For comparison, we also tested the efficiency of EC-guided adaptive sampling[42,64], where in each round, 10 states with the minimal EC residue pair distances were picked for adaptive sampling.

MoaD-MoaE association. 55 μs of previously published implicit solvent MD simulations on MoaD and MoaE association were used in this study[42]. The trajectories were clustered into 500 states using k-means algorithm (see Supplementary Information for details). An MSM was constructed with a lag time τ of 40 ns chosen based on convergence of the implied timescales (Supplementary Fig. S1d). The SAXS profile of the native MoaD-MoaE complex was calculated from 100 snapshots of 50 ns implicit solvent MD simulation on the MoaD-MoaE complex crystal structure (PDB ID: 1FM0[65]), which was also taken from the previous study[42]. At each scattering verctor q_i, the average of scattering intensities ($I(q_i)$) and the standard deviation of $I(q_i)$ were determined as $I_{target}(q_i)$ and $\sigma_{target}(q_i)$, respectively. 100 structures from each state were randomly extracted to calculate the SAXS profiles of each state. The SAXS discrepancy values between the SAXS profiles of each state and the target were calculated using the reduced χ^2 function. The average SAXS discrepancy value of each state was calculated from the lowest 50 discrepancy values, in order to reduce the statistical errors due to clustering.

The distances of five evolutionarily coupled MoaD-MoaE residue pairs (E12-R127, R11-E53, A54-M58, Q57-K61, T58-K61) with the highest EVcomplex scores determined in the previous study[66] were also calculated from the 50 structures with the lowest SAXS discrepancy scores from each state. Different kinetic MC sampling

algorithms were carried out following the same protocol as in the folding systems. The initial state was the state with the largest average residue pair distance and the predicted dimeric state was chosen as the final state. The sampling times required to reach the predicted dimeric state were calculated to quantify the sampling efficiency using different protocols.

Homodimeric PYR1 association. SAXS-guided MD simulations were performed to predict the dimeric PYR1 structure. We performed 10 ns explicit solvent MD simulations on the PYR1 crystal structure (PDB ID: 3K3K[43]) with and without position restraints on protein heavy atoms. The resulting trajectories (each with 1000 frames) were given as the input of the WAXSiS to calculate their theoretical SAXS profiles. The errors of the calculated SAXS profiles were determined by the WAXSiS automatically. The experimental SAXS data adapted from Nishimura *et al.*[43] (Bioisis ID: BID_1PYR1P, http://www.bioisis.net/experiments/44) were fitted to the calculated SAXS profiles by minimizing the following χ^2 function (on logarithmic scale, equation 3) as implemented in WAXSiS[51,52]:

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\left[logI_{cal}(q_i) - log(fI_{exp}(q_i) + c)\right]^2 \quad (3)$$

where N is the total number of data points, q_i is the momentum transfer, $I_{cal}(q_i)$ and $I_{exp}(q_i)$ are the calculated and experimental scattering intensity at q_i, f and c are the fitting parameters. We find that the experimental SAXS data fit better to the SAXS profile calculated from free MD simulations with a χ^2 of 0.006 (as compared to 0.010 for the SAXS profile calculated from constraint MD simulations, Supplementary Fig. S3). The SAXS profile calculated from free MD simulations was used for the final structure identification.

Initially, two monomers of the crystal structure were separated with their center of mass distance approximately at 50 Å in VMD[67]. The structures were solvated with TIP3P water and Na^+/Cl^- were added to the system with the salt concentration of 0.15 M using AmberTools15. The simulations were performed in Amber14 using the Amber ff14SB force field[58]. The simulation protocol and parameters were the same as described in the MD simulations of experimentally determined structures of single domain proteins. A single simulation was started from the equilibrated structure and run for 100 ns and the resulting trajectory was clustered into 100 states using a similar clustering scheme as in the MoaD-MoaE system. The SAXS profiles of all states were calculated. 25 clusters with the lowest SAXS discrepancies (calculated using χ^2 function on logarithmic scale) from the target SAXS profile were chosen as the seeding structures for the second round of sampling. In the second round, each trajectory was run for 100 ns. The trajectories were clustered into 200 states, and again 25 clusters with the lowest SAXS discrepancy scores were chosen for the third round of sampling. In the third round, each trajectory was run for 60 ns, yielding total simulation time of around 4 μs. Trajectories in the last two rounds were clustered into 200 states for final analysis. The SAXS profiles of these 200 states were calculated and the SAXS discrepancy scores were computed using the reduced χ^2 function for consistency. In order to further improve the accuracy of the structural model obtained from the three rounds of adaptive sampling, 5 parallel simulations were run for 20 ns each starting from the closest near-dimeric state among the 200 states.

Results

SAXS along with thermodynamic information is sufficient to distinguish the native folds of small proteins. As an illustration of utilizing SAXS information to predict the native state structure of proteins, we first studied the folding of HP35 double norleucine mutant, protein G and α3D domain. The three systems with varying number of residues (35, 56, 73) were chosen to explore how the prediction accuracy changes as protein size increases due to the inherent ambiguity of SAXS data. MSMs were constructed for each system from the extensive amount of folding trajectories. The SAXS profiles calculated from short MD simulations on the experimentally determined structures or the structure of the closest homolog in the PDB (PDB IDs: 2F4K[54], 1MI0[56], 2A3D[57]) were used to obtain the target SAXS profiles for structure prediction. The SAXS discrepancy scores (reduced χ^2) between the SAXS profile of each state and the target SAXS profile were computed. To compare the SAXS discrepancy scores of all states, we can predict the near-folded state and further test whether SAXS is sufficient to make good predictions by aligning the predicted structure to the experimentally determined structure.

This is shown by plotting the free energies of all MSM states with respect to their SAXS discrepancy scores, as shown in Fig. 1a–c. Without any *a priori* knowledge, the free energy information estimated from the MSMs can serve as additional metric to identify the near-native or intermediate states of proteins from simulation. Combining both free energy and SAXS discrepancy information, we seek to make prediction of the native folds of proteins. For these three folding systems, generally, high free energy states tend to have much higher SAXS discrepancy values, while low free energy states corresponding to more stable near-native or intermediate structures tend to have much lower SAXS discrepancy values (Fig. 1a–c). The RMSD plots in Supplementary Fig. S4 also suggest that the states with high SAXS discrepancy scores have large C_α RMSDs from their crystal structures, and *vice versa*. Though in HP35 a few states with minimal SAXS discrepancy scores have high free energy (Fig. 1a), these structures are actually close to the native structures (Supplementary Fig. S4). However, as protein size increases, there are more low free energy states (<1 kcal/mol) with comparably low discrepancy values in α3D as compared to the states in smaller HP35 and protein G, which will be hard to distinguish using SAXS information alone. This is consistent with the underdetermined nature of SAXS pattern. All together, this suggests that for the folding of small single domain proteins, thermodynamic information estimated from the MSMs is valuable for dealing with the ambiguity of SAXS data.

Based on both free energy and SAXS discrepancy information, the states with the lowest free energy were predicted as the near folded states from the simulation datasets. The SAXS profiles of the target and the predicted

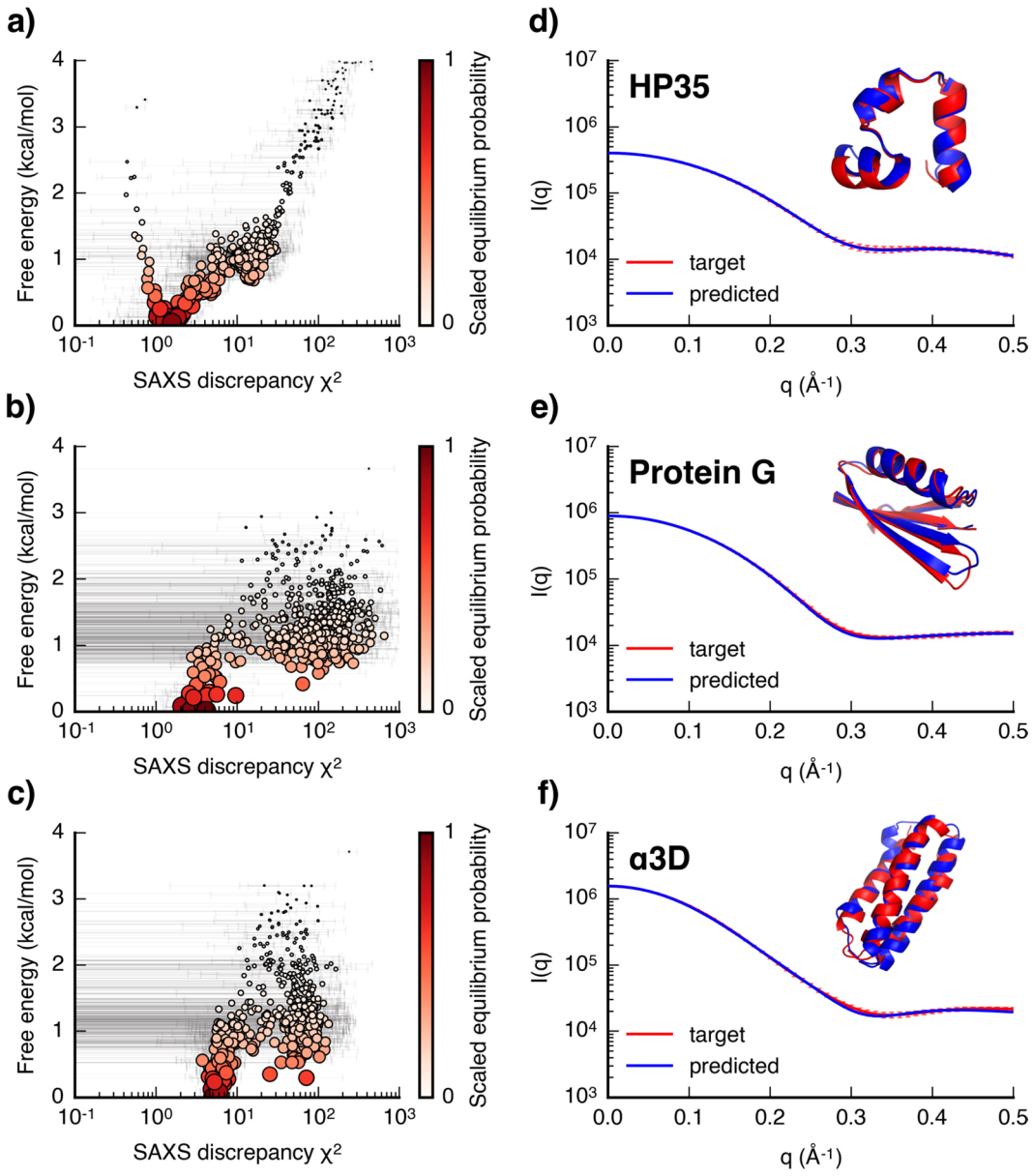

Figure 1. Predicting native folds of single domain proteins. The plots of individual MSM state free energy with respect to their average SAXS discrepancy values (χ^2) for the (**a**) HP35, (**b**) protein G, and (**c**) α3D. Each dot represents a MSM state. The dot size and its color darkness are scaled by the equilibrium probability of that state estimated from the MSM. The errorbars for the SAXS discrepancy values of all states are shown in grey line. SAXS profiles of the target (red) and the predicted (blue) states, and overlays of the crystal structure (red) and the simulation predicted structure (blue) for (**d**) HP35, (**e**) protein G, and (**f**) α3D. The errorbars for the target SAXS profiles are also shown in the figure.

folded states for the three proteins are shown in Fig. 1d–f. The predicted post-MD model SAXS profiles of HP35, protein G and α3D fit to their target SAXS profiles with reduce χ^2 values of 0.045, 0.775 and 0.975. The residuals plots suggest that the discrepancies between the predicted and the target SAXS profiles are comparable to the errors on the target SAXS profile (Supplementary Fig. S5). The fittings were also assessed using the correlation map method[68], which also suggest high similarities between the target and predicted SAXS profiles (Supplementary Fig. S6). Overlays of the native structures and the predicted folded structures of HP35, protein G and α3D give the C_α RMSDs of 0.7, 0.78 and 2.75 Å respectively. Radius of gyration (R_g) and maximum diameter (D_{max}) estimated from the target and predicted SAXS profiles using Guinier analysis are in good agreement (Table 1).

An accurate estimate of free energy values from the MSMs usually requires sufficient amount of sampling. To further improve the accuracy of identifying native structures from simulation datasets, other types of structural information can be incorporated together with SAXS data. For example, we demonstrate that in combination with the distance restraints inferred from a few top ranked evolutionarily coupled residue pairs, the near-native

Systems	R_g (Å)		D_{max} (Å)	
	native	predicted	native	predicted
HP35	10.96	10.90	35.61	33.29
Protein G	12.25 ± 0.01	12.27 ± 0.01	37.04	39.76
α3D	14.34 ± 0.01	14.61 ± 0.02	45.58	45.32
MoaD-MoaE	21.38	21.55 ± 0.03	72.67	76.07
PYR1	23.89 ± 0.08 (calc.) 23.72 ± 0.6 (expr.)	23.92 ± 0.07	67.4 (calc.) 68.46 (expr.)	67.57

Table 1. Comparisons of the radius of gyration (R_g) and the maximum diameter (D_{max}) of the native and the predicted states estimated from the SAXS data.

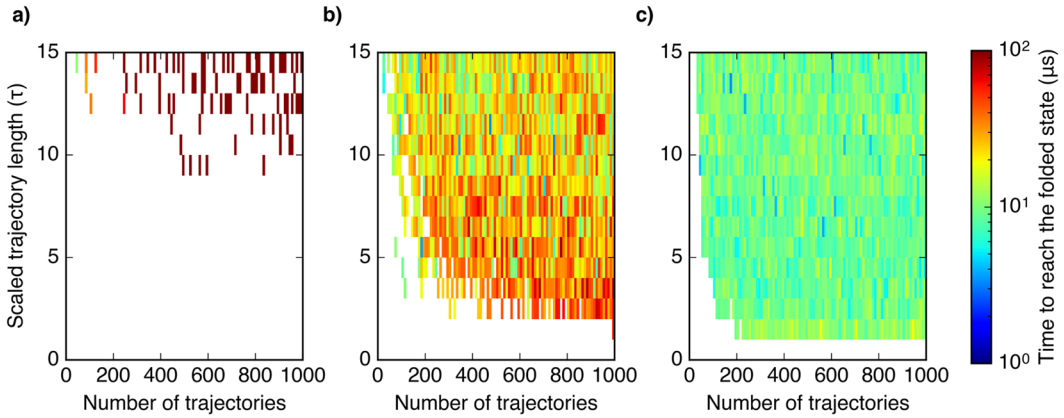

Figure 2. Enhanced efficiency in sampling the folding of HP35. Total simulation time required to reach the folded state from an arbitrary unfolded state for sets of samplings using (**a**) traditional long simulation, (**b**) random adaptive sampling, and (**c**) SAXS-guided adaptive sampling. Scaled trajectory length is the length of each individual trajectory in each specific sampling scheme by the lag time τ of the MSM. Number of trajectories is the total number of trajectories run for each sampling scheme, given by the product of the number of parallel trajectories and the number of sampling rounds. The average total required sampling times using the 3 different protocols over 1500 sets of samplings (excluding the sets of sampling that do not reach the target native state) are 235.03 μs, 27.61 μs, 9.76 μs.

states of HP35 and protein G can be identified (Supplementary Fig. S2). This hybrid information could tackle with the challenge of SAXS inherent ambiguity and possible thermodynamic inaccuracy due to insufficient sampling.

Enhanced efficiency in sampling protein folding. To test the efficiency of utilizing the SAXS discrepancy information in sampling protein folding, we employed kinetic MC simulations of the folding of HP35, protein G and α3D on the MSMs using different sampling strategies. The total sampling times required for transition from an arbitrary expanded unfolded state to the predicted folded state were calculated to compare the overall sampling efficiency of different sampling strategies. Figure 2 shows the results for sampling the folding of HP35 using traditional long simulation, random and SAXS-guided adaptive samplings. It is clearly shown in Fig. 2a that traditional way of running long simulations takes the longest time to discover the folded state. Few simulation sets with individual trajectory length shorter than 10 τ (300 ns MD simulation) can reach the folded state even with 1000 trajectories running in parallel. Adaptive sampling can effectively reduce the simulation time to reach the folded state. With random adaptive sampling (Fig. 2b), namely randomly picking seeds for iterative sampling, an order of magnitude decrease of computational time to reach the folded state over traditional long simulation is observed. In addition, in the short individual trajectory length regions where traditional long simulation sets can never reach the folded state, random adaptive sampling can reach the folded state in tens of microseconds. Figure 2c shows SAXS-guided adaptive sampling can even further decrease the sampling time than random adaptive sampling. The total simulation time required for obtaining the native state is $\sim$10 μs. Enhanced efficiency in SAXS-guided adaptive sampling is also observed in sampling protein G and α3D (Supplementary Figs. S7 and S8). Using another metric, the number of MSM states explored using random and SAXS-guided adaptive sampling, to compare their sampling efficiency, it is clearly shown that SAXS-guided sampling enhances sampling efficiency by reducing the sampling of 'insignificant' states (Supplementary Fig. S9), which have large deviations from the target as measured by SAXS-discrepancy scores. Overall, these results suggest that utilizing the SAXS information in adaptive sampling can effectively reduce computational time required to discover the folded structures of small proteins and the folding pathways in unbiased simulations. We further tested the efficiency of utilizing both SAXS and distance restraints inferred from ECs in adaptive sampling of protein folding of HP35 and protein G (Supplementary Figs. S7 and S10). As compared to SAXS-guided or EC-guided adaptive sampling, we

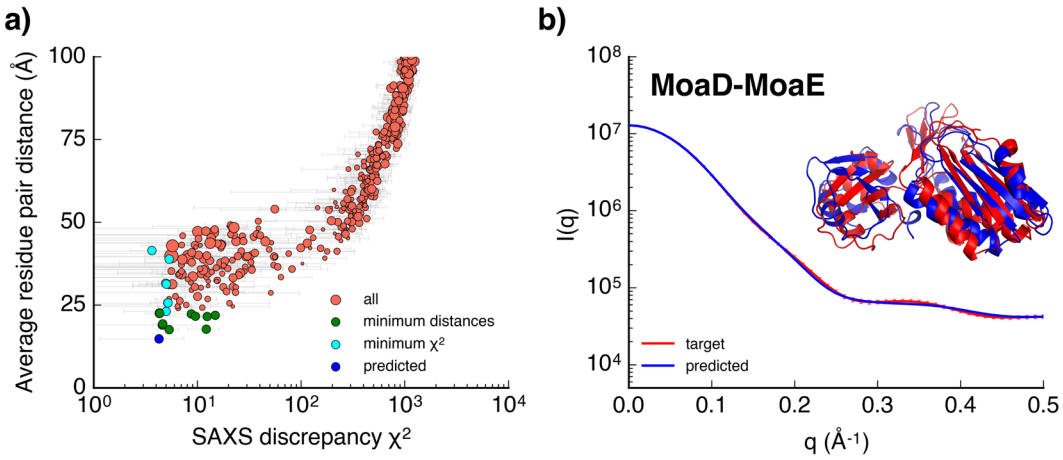

Figure 3. Predicting the structure of MoaD-MoaE complex. (**a**) The plots of the average C_α distance of five evolutionarily coupled residue pairs of each MSM state with respect to their average SAXS discrepancy scores (χ^2). Each dot represents an MSM state. The dot size is scaled by the equilibrium probability of that state estimated from the MSM. The 10 states with minimal average residue pair distances and the 10 states with minimal χ^2 are colored in green and cyan, respectively (with 3 overlapped states colored in green and the overlapped predicted state colored in blue). (**b**) SAXS profiles of the target (red) and the predicted (blue) states. The errorbars for the target SAXS profile are also shown in the figure. Overlay of the crystal structure (red) and the simulation predicted structure (blue) gives a C_α RMSD of 5.32 Å from the crystal structure.

further observe a slight enhancement of sampling efficiency using this hybrid approach (Supplementary Figs. S7 and S10).

SAXS along with distance restraints predict the near-crystal structure of MoaD-MoaE complex. The association of *E. coli* molybdopterin synthase subunits, MoaD and MoaE, was used to explore the application of SAXS-guided adaptive sampling approach in predicting dimeric protein structures and association pathways. As in the folding systems, the SAXS profiles of all MSM states were calculated, and the SAXS profile calculated from short MD simulations of the crystal structure (PDB ID: 1FM0)[65] was used as the target SAXS profile. We first tested whether SAXS in combination with thermodynamic information is sufficient to make good predictions of dimeric structure from MD simulation datasets by comparing the SAXS profiles of each state and the target. Supplementary Fig. S11 gives the plots of free energies of all states with respect to their SAXS discrepancy scores. Unlike the free energy plots in folding systems, there is not a clear correlation between the free energy of each state and its SAXS discrepancy value. Although several most populated states estimated by the MSM have relatively low SAXS discrepancy values, there are many less populated states that can give even lower SAXS discrepancy values, which might imply their higher structure similarity to the target structure. From this, we speculate that the most populated state might not be the actual near-dimer structure but a thermodynamically metastable state, which is possible considering the insufficient sampling of the protein-protein association ensembles.

This prompts us to look for additional information to further distinguish the states displaying similar SAXS profiles. We explored to combine the distance restraints inferred from intermolecular ECs with SAXS data to improve structure prediction accuracy. Intermolecular ECs can provide valuable insights into residue contacts at protein-protein interface[66]. Top 5 five ranked EC residue pairs were chosen, and the distances between these residue pairs for all states were calculated. We plotted the average residue pair distances of all states with respect to their SAXS discrepancy scores to characterize their structure differences (Fig. 3a). We observe that there is still an overall correlation between the average residue pair distance and the SAXS discrepancy scores, though states with approximately equal SAXS discrepancy scores can have large varying average residue pair distances. For example, the 10 states with the minimal SAXS discrepancy scores (cyan, Fig. 3a) have significant differences in interfacial residue pair distances (range between 15–40 Å). As shown in Supplementary Fig. S12, these states all display similar SAXS profiles as compared to the target SAXS data, however, the MoaD-MoaE complexes adopt completely different orientations. The inherent ambiguity of SAXS data is much more obvious for MoaD-MoaE complex as compared to smaller single domain proteins. Nevertheless, incorporating distance restraints information at the complex interface effectively distinguish the states that display equally consistent SAXS profiles as the target SAXS profile.

Integrating both SAXS discrepancy scores and distance restraints information, we predict the state with the lowest interfacial residue pair distance, which also has relatively low SAXS discrepancy score, as the near-dimer structure (blue, Fig. 3a). Figure 3b shows the predicted structure aligns well with the crystal structure (C_α RMSD: 5.32 Å) and the SAXS profile of the predicted structure also matches well with the target SAXS profile (reduced $\chi^2 = 0.922$, residuals plots shown in Supplementary Fig. S13). Fitting assessed using the correlation map method also suggests that the predicted SAXS profile adequately describes the target SAXS data (Supplementary Fig. S14). The relatively large RMSD could be due to the loop of MoaD that is inserted into MoaE to form the active site in

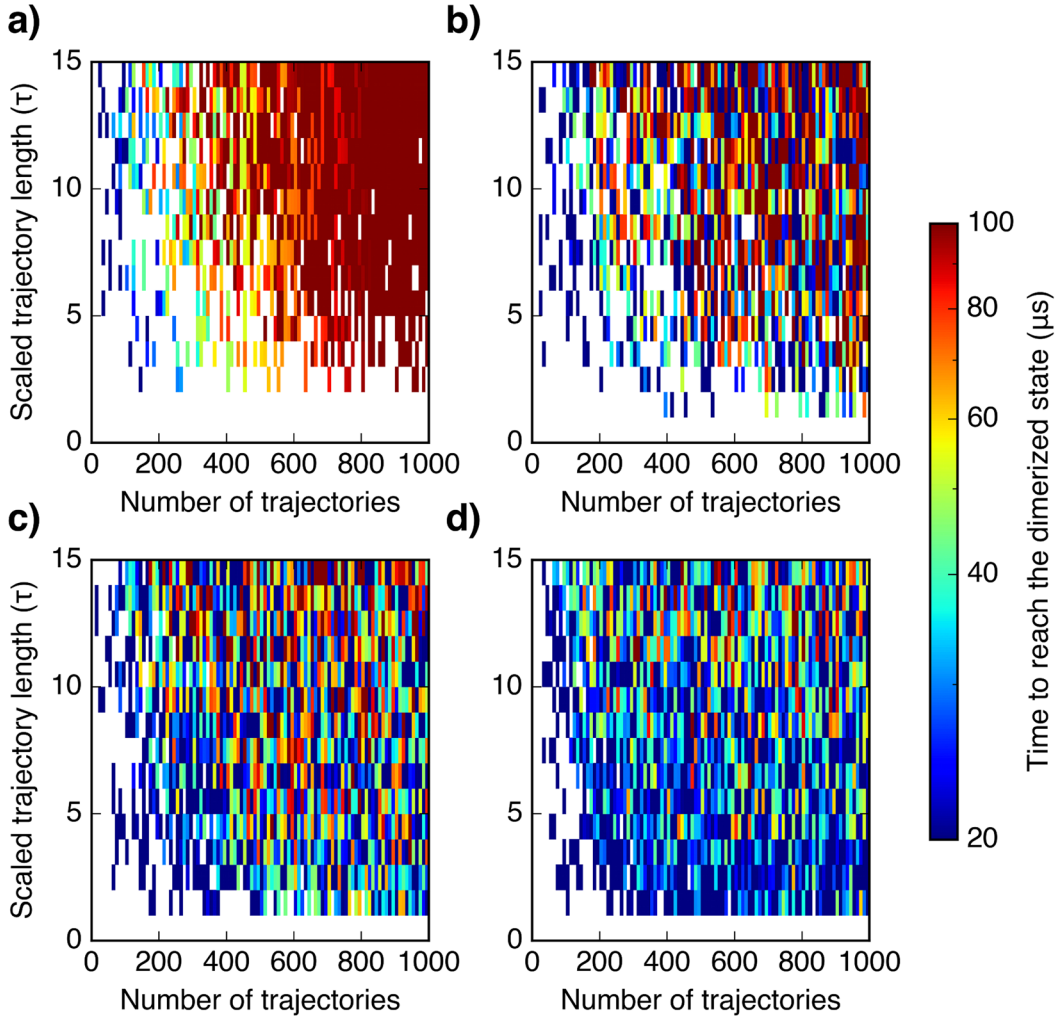

Figure 4. Enhanced efficiency in sampling the association of MoaD-MoaE. Total simulation time required to reach the predicted dimeric state from an arbitrary unassociated state for sets of samplings using (**a**) traditional long simulation, (**b**) random adaptive sampling, (**c**) SAXS-guided adaptive sampling and (**d**) SAXS-EC-guided adaptive sampling. Scaled trajectory length is the length of each individual trajectory in each specific sampling scheme by the lag time τ of the MSM. Number of trajectories is the total number of trajectories run for each sampling scheme, given by the product of the number of parallel trajectories and the number of sampling rounds. The average total required sampling times using the 4 different protocols over 1500 sets of samplings (excluding the sets of samplings that do not reach the target state) are 113.28 μs, 61.16 μs, 41.57 μs, and 30.64 μs.

the crystal structure[65]. This process has not been captured from the dataset[42]. The R_g and D_{max} values estimated from the target and predicted SAXS profile are also in good agreement (Table 1).

Supplementary Fig. S15 shows the comparison of the SAXS profiles and the snapshots of the 10 states with the minimal average residue pair distances (green, Fig. 3a). As compared to the predicted state, the states with relatively higher SAXS discrepancy scores show larger RMSDs from the crystal structure. A combination of SAXS and interfacial residue contact information gives the best structure prediction. All together, these results demonstrate that a hybrid approach that combines SAXS with distance restraints information provides a good structure prediction of protein-protein complex.

Enhanced efficiency in sampling protein-protein association. In order to test the feasibility of utilizing SAXS to accelerate unbiased sampling of protein association pathways, we performed kinetic MC samplings on the MoaD-MoaE MSM using different protocols, including traditional long simulation, random adaptive sampling, SAXS-guided as well as SAXS-EC guided adaptive samplings. We calculated the total sampling time required to observe the transition from an arbitrarily chosen unassociated state to the predicted near-dimeric state, as shown in Fig. 4. As in the folding systems, adaptive sampling strategy effectively reduces the total sampling times as compared to long serial simulations. As compared to random adaptive sampling, SAXS-guided adaptive sampling also improves the sampling efficiency. In the 1500 sets of samplings, the average total sampling time to reach the target state is ~60 μs using random adaptive sampling, and the required sampling time

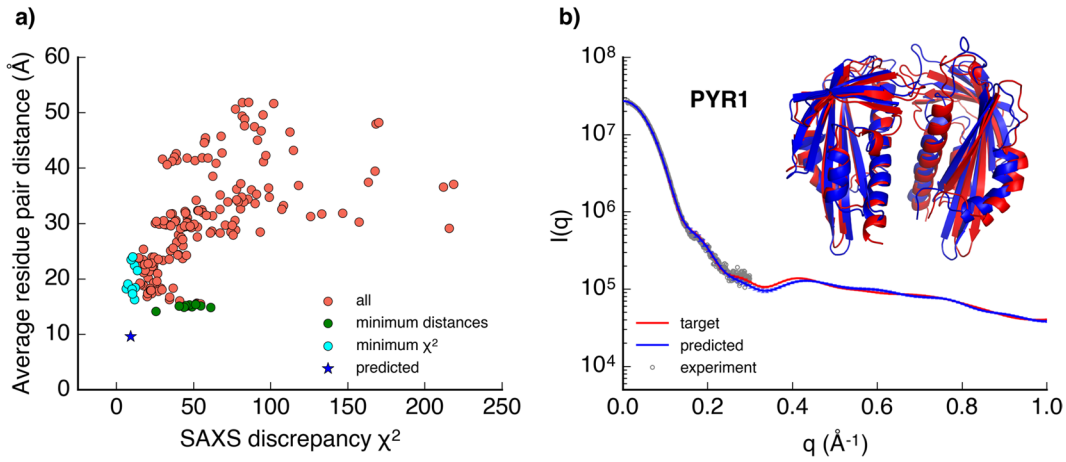

Figure 5. Predicting the PYR1 structure. (**a**) The plot of the average C_α distance of two residue pairs of individual state with respect to the SAXS discrepancy value χ^2. Each circle represents a single state from the clustering. The 10 states with the minimal average residue pair distances are colored in green, and the 10 states with the minimal χ^2 are colored in cyan, respectively. The blue star denotes the refined predicted PYR1 structure obtained after 20 ns MD simulation starting from the state with the minimal residue pair distance. (**b**) SAXS profiles of the target (red) and the predicted (blue) states are shown with the errorbars. Overlay of the crystal structure (red) and the best simulation predicted structure (blue) gives a C_α RMSD of 3.18 Å. The fitted SAXS experimental data are marked in grey circles.

decreases to ~40 μs using SAXS-guided sampling. In the case of hybrid approach using SAXS-EC-guided adaptive sampling, we observe a further improvement of the sampling efficiency, and the average required sampling time is ~30 μs. When compared with the performances of SAXS-guided or EC-guided adaptive sampling, SAXS-EC-guided adaptive sampling performs better as compared to both sampling strategies (Supplementary Fig. S16). All together, these results suggest that SAXS-guided sampling approach enhances time efficiency in sampling protein-protein association, and by combining both SAXS and distance restraints, the sampling efficiency is further enhanced.

SAXS-guided adaptive sampling provides a near-crystal structure PYR1 complex. The last example is to test the efficiency of using the SAXS-guided adaptive sampling in actual MD simulations to predict the complex structure of PYR1. The target SAXS data was computed from short explicit-solvent MD simulations on the PYR1 crystal structure (PDB ID: 3K3K)[43]. After each round of sampling, all protein conformations were clustered and the SAXS discrepancy scores of all states were calculated and used in the adaptive sampling. By the third round of sampling, the structure with C_α RMSD from the crystal structure of 5.14 Å was achieved, with total sampling time of 4 μs. The trajectories in last two rounds were clustered into 200 states, and the SAXS profiles of these 200 states were computed and the SAXS discrepancy scores between the target and each state were calculated using the reduced χ^2 function. Two pairs of residues from each monomer (K63-D155, L166-L166) at the interface of the crystal structure were chosen to calculate the distances to characterize the structural similarity to the crystal structure.

Figure 5a gives the plots of average residue pair distances of all states with respect to their SAXS discrepancy values. Similar to the association of MoaD and MoaE, even at low SAXS discrepancy region, there are multiple states that have approximately equal SAXS discrepancy values but varying residue pair distance; while the structures with approximately equal average residue pair distances can have varying SAXS discrepancy values. Supplementary Fig. S17 shows the SAXS profiles and the snapshots of the 10 states with the minimal SAXS discrepancy scores. Despite high similarities between the SAXS profiles of these states as compared to the target SAXS profile, the two monomers adopt various types of orientations and have varying degrees of deviations from the crystal structure. These results again highlight the inherent ambiguity of SAXS patterns. Supplementary Fig. S18 shows the SAXS profiles and the snapshots of the 10 states with the minimal interfacial residue pair distances. From the three rounds of adaptive samplings, the state with the lowest interfacial residue pair distance and a relatively small SAXS discrepancy score (reduced $\chi^2 = 25.561$) gives the closest structural model to the PYR1 crystal structure (C_α RMSD: 5.14 Å, Supplementary Fig. S18b).

To further improve the quality of the PYR1 structural model, we performed additional MD simulations to refine the structural model obtained from the three rounds of adaptive sampling. Briefly, 5 parallel simulations were launched from the state with the minimal residue pair distance (among the 200 states) and run for 20 ns each. The last 10 ns simulation data from each trajectory were used to calculate the SAXS profiles. The conformation from the trajectory with the lowest reduced χ^2 was determined as the predicted PYR1 structural model (Fig. 5a, blue star). The reduced χ^2 between the predicted and the target SAXS profiles decreases to 9.111 and the C_α RMSD of the predicted structural model from the crystal structure decreases to 3.18 Å (Fig. 5b). The target SAXS profile, and the SAXS profile of the predicted structure are shown in Fig. 5b. The target SAXS profile matches well with the previously published SAXS experimental data[43] within $q < 0.3$ Å^{-1} region. The R_g and D_{max} values estimated from the experimental SAXS data and the target and predicted SAXS profiles are compared in

Table 1. The residuals plots suggest that the major discrepancies between the predicted and the target SAXS profiles are from the errors of the predicted SAXS profiles (Supplementary Fig. S19). We believe more sampling from the predicted structural model will further improve the structural prediction accuracy and the corresponding SAXS profile will have even better match with the target SAXS profile. Overall, these results demonstrate that SAXS-guided adaptive sampling is an efficient sampling approach to predict protein complex structures from unbiased all atom MD simulations.

Discussion

Long timescale unbiased MD simulations can be a complementary method to fully interpret the limited structural information contained in SAXS data, and predict accurate protein structures, ensembles and dynamics. In this study, we have demonstrated the utility of SAXS and hybrid information in adaptive sampling process to enhance time efficiency in unbiased sampling of protein folding and protein-protein association pathways. By analyzing the extensive protein folding and protein-protein association simulation datasets, we demonstrate the use of SAXS data along with thermodynamics or distance restraints information in improving the accuracy of structure prediction from MD simulations. For the folding of small proteins, we show that SAXS data in combination with free energy information estimated from the MSMs is sufficient to distinguish the native states from simulation datasets (Fig. 1). Distance restraints which can be inferred from intramolecular EC can be combined with SAXS data to further distinguish the internal structure differences of conformational states that display similar SAXS profiles. For the association of MoaD-MoaE (Fig. 3) and homodimer PYR1 (Fig. 5), integrating SAXS data and interfacial distance restraints, good predictions of near-native complex structures can be obtained using this approach. For practical applications, which additional external information may be required for further structure differentiations can also be obtained from these structural models predicted from the simulation. Based on this prior knowledge, relevant computations or experiments could be performed to provide the information[69–71].

From kinetic MC sampling, we have shown that the computational times in sampling either protein folding (Fig. 2, Supplementary Figs. S7, S8) or protein-protein association (Fig. 4) are significantly reduced by incorporating SAXS and SAXS-based hybrid information as reaction coordinates in adaptive sampling. Furthermore, by combining both SAXS and distance restraints in adaptive sampling, the sampling efficiency is better than the sampling guided by either type of structural information alone. We expect that these hybrid approaches will be useful for the study of larger proteins, as the inherent ambiguity of SAXS data would be more significant. It should be noted that these approaches provide not only the final predictions of native states of proteins and complexes, but also structure ensembles and dynamics. During the sampling process, our protocol bias sampling directions towards the target to prevent exploring 'irrelevant' states as defined according to the adaptive seed selection criteria. After the target structure is discovered, one can do more sampling along the initial sampled pathways to collect accurate structural ensembles and dynamics. For heterogeneous ensembles, the obtained ensembles could be reweighted against experimental data to be further refined.

Our approach has some similarities to a previously proposed experiment-guided sampling technique, PaCS-Fit[72]. PaCS-Fit also involves iteratively running short parallel simulations and picking conformations that are 'closer' to experimental data for further sampling except the clustering step as implemented in our MSM-based adaptive sampling strategy. The clustering step is essential for two reasons. First, due to SAXS underdetermined nature, sturctures with similar molecular envelopes may display similar SAXS profiles but can have essential structure differences at atomic resolution. This is already demonstrated in our MoaD-MoaE and PYR1 systems. Without clustering, in each cycle PaCS-Fit will likely only select redundant conformations that are structurally similar to continue sampling, and therefore bias simulations to an ensemble consistent with target experimental data but structurally distant from target structure. Instead, by iteratively running simulations in parallel from multiple cluster states, we can likely achieve the structure ensemble reasonably 'close' to the true ensemble. Second, the clustering can effectively reduce the amount of SAXS calculations required for long timescale simulations. Peng *et al.*[72] has demonstrated the success of application of PaCS-Fit in predicting small conformational changes of proteins. Our study has addressed some challenges as we extend the application of this approach in studying the protein folding and protein-protein association process.

Finally, using MSMs, the large amount of unbiased simulation data collected using adaptive sampling approaches could be merged to construct discrete protein dynamic models. Using transition path theory[73,74], protein dynamics such as protein folding and protein-protein pathways can be fully mapped out from the MSMs. All together, SAXS-based adaptive sampling along with MSMs potentially allows us to predict not only the protein functional conformations but also the pathways of conformational changes with a reasonable computational cost. This method can be useful in determining the unknown structures of proteins and complexes. At the same time, the constructed models from simulation can be updated when more accurate or orthogonal experimental information is available[69,70]. Other experimental information can be incorporated in a similar manner to collectively enhance the sampling and improve the accuracy of prediction from the simulation. These methods provide a way to build a dynamic model of protein function consistent with the available experimental and computational data.

References

1. Spreitzer, R. J. & Salvucci, M. E. Rubisco: structure, regulatory interactions, and possibilities for a better enzyme. *Annu. Rev. Plant Biol.* **53**, 449–475 (2002).
2. Smalle, J. & Vierstra, R. D. The ubiquitin 26S proteasome proteolytic pathway. *Annu. Rev. Plant Biol.* **55**, 555–590 (2004).
3. Moffett, A. S. & Shukla, D. Using molecular simulation to explore the nanoscale dynamics of the plant kinome. *Biochem. J.* **475**, 905–921 (2018).
4. Moffett, A. S., Bender, K. W., Huber, S. C. & Shukla, D. Allosteric control of a plant receptor kinase through sglutathionylation. *Biophys. J.* **113**, 2354–2363 (2017).
5. Moffett, A. S., Bender, K. W., Huber, S. C. & Shukla, D. Molecular dynamics simulations reveal the conformational dynamics of Arabidopsis thaliana bri1 and bak1 receptor-like kinases. *J. Biol. Chem.* **292**, 12643–12652 (2017).

6. Tan, X. *et al*. Mechanism of auxin perception by the TIR1 ubiquitin ligase. *Nature* **446**, 640 (2007).
7. Melcher, K. *et al*. A gate–latch–lock mechanism for hormone signaling by abscisic acid receptors. *Nature* **462**, 602 (2009).
8. Murase, K., Hirano, Y., Sun, T.-P. & Hakoshima, T. Gibberellin-induced DELLA recognition by the gibberellin receptor GID1. *Nature* **456**, 459 (2008).
9. Vanatta, D. K., Shukla, D., Lawrenz, M. & Pande, V. S. A network of molecular switches controls the activation of the two-component response regulator ntrc. *Nat. Commun.* **6**, 7283 (2015).
10. Shukla, D., Peck, A. & Pande, V. S. Conformational heterogeneity of the calmodulin binding interface. *Nat. Commun.* **7**, 10910 (2016).
11. Selvam, B., Shamsi, Z. & Shukla, D. Universality of the sodium ion binding mechanism in class a g-proteincoupled receptors. *Angew. Chem. Int. Ed.* **130**, 3102–3107 (2018).
12. Meng, Y., Shukla, D., Pande, V. S. & Roux, B. Transition path theory analysis of c-src kinase activation. *Proc. Natl. Acad. Sci. USA* **113**, 9193–9198 (2016).
13. Mertens, H. D. & Svergun, D. I. Structural characterization of proteins and complexes using small-angle X-ray solution scattering. *J. Struct. Biol.* **172**, 128–141 (2010).
14. Deniz, A. A. Deciphering complexity in molecular biophysics with single-molecule resolution. *J. Mol. Biol.* **428**, 301–307 (2016).
15. Jeschke, G. Deer distance measurements on proteins. *Annu. Rev. Phys. Chem.* **63**, 419–446 (2012).
16. Allison, J. R. Using simulation to interpret experimental data in terms of protein conformational ensembles. *Curr. Opin. Struct. Biol.* **43**, 79–87 (2017).
17. Hub, J. S. Interpreting solution x-ray scattering data using molecular simulations. *Curr. Opin. Struct. Biol.* **49**, 18–26 (2018).
18. Chen, Y., Campbell, S. L. & Dokholyan, N. V. Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophys. J.* **93**, 2300–2306 (2007).
19. Grubisic, I., Shokhirev, M. N., Orzechowski, M., Miyashita, O. & Tama, F. Biased coarse-grained molecular dynamics simulation approach for flexible fitting of X-ray structure into cryo electron microscopy maps. *J. Struct. Biol.* **169**, 95–105 (2010).
20. Yang, S., Blachowicz, L., Makowski, L. & Roux, B. Multidomain assembled states of Hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci. USA* **107**, 15757–15762 (2010).
21. Zhu, G., Saw, W. G., Nalaparaju, A., Grüber, 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* **121**, 2252–2264 (2017).
22. Shi, J. *et al*. Atomistic structural ensemble refinement reveals non-native structure stabilizes a sub-millisecond folding intermediate of CheY. *Sci. Rep* **7**, 44116 (2017).
23. Shevchuk, R. & Hub, J. S. Bayesian refinement of protein structures and ensembles against saxs data using molecular dynamics. *plos Comput. Biol.* **13**, e1005800 (2017).
24. Pitera, J. W. & Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* **8**, 3445–3451 (2012).
25. Vashisth, H., Skiniotis, G. & Brooks, C. L. III. Using enhanced sampling and structural restraints to refine atomic structures into low-resolution electron microscopy maps. *Structure* **20**, 1453–1462 (2012).
26. Björling, A., Niebling, S., Marcellini, M., van der Spoel, D. & Westenhoff, S. Deciphering solution scattering data with experimentally guided molecular dynamics simulations. *J. Chem. Theory Comput.* **11**, 780–787 (2015).
27. Orzechowski, M. & Tama, F. Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys. J.* **95**, 5692–5705 (2008).
28. Granata, D., Camilloni, C., Vendruscolo, M. & Laio, A. Characterization of the free-energy landscapes of proteins by NMR-guided metadynamics. *Proc. Natl. Acad. Sci. USA* **110**, 6817–6822 (2013).
29. White, A. D., Dama, J. F. & Voth, G. A. Designing free energy surfaces that match experimental data with metadynamics. *J. Chem. Theory Comput.* **11**, 2451–2460 (2015).
30. Kimanius, D., Pettersson, I., Schluckebier, G., Lindahl, E. & Andersson, M. SAXS-guided metadynamics. *J. Chem. Theory Comput.* **11**, 3491–3498 (2015).
31. Yang, S. Methods for SAXS-Based Structure Determination of Biomolecular Complexes. *Adv. Mater.* **26**, 7902–7910 (2014).
32. Cordeiro, T. N. *et al*. Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Curr. Opin. Struct. Biol.* **42**, 15–23 (2017).
33. Konarev, P. V. & Svergun, D. I. *A posteriori* determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. *IUCrJ* **2**, 352–360 (2015).
34. Petoukhov, M. V. & Svergun, D. I. Ambiguity assessment of small-angle scattering curves from monodisperse systems. *Acta Crystallogr. D Biol. Crystallogr* **71**, 1051–1058 (2015).
35. Boldon, L., Laliberte, F. & Liu, L. Review of the fundamental theories behind small angle x-ray scattering, molecular dynamics simulations, and relevant integrated application. *Nano Rev* **6**, 25661 (2015).
36. Huang, X., Bowman, G. R., Bacallado, S. & Pande, V. S. Rapid equilibrium sampling initiated from nonequilibrium data. *Proc. Natl. Acad. Sci. USA* **106**, 19765–19769 (2009).
37. Beauchamp, K. A., Ensign, D. L., Das, R. & Pande, V. S. Quantitative comparison of villin headpiece subdomain simulations and triplet–triplet energy transfer experiments. *Proc. Natl. Acad. Sci. USA* **108**, 12734–12739 (2011).
38. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
39. Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014).
40. Shukla, D., Hernández, C. X., Weber, J. K. & Pande, V. S. Markov state models provide insights into dynamic modulation of protein function. *Acc. Chem. Res.* **48**, 414–422 (2015).
41. Metzner, P., Noé, F. & Schütte, C. Estimating the sampling error: Distribution of transition matrices and functions of transition matrices for given trajectory data. *Phys. Rev. E* **80**, 021106 (2009).
42. Shamsi, Z., Moffett, A. S. & Shukla, D. Enhanced unbiased sampling of protein dynamics using evolutionary coupling information. *Sci. Rep* **7**, 12700 (2017).
43. Nishimura, N. *et al*. Structural mechanism of abscisic acid binding and signaling by dimeric PYR1. *Science* **326**, 1373–1379 (2009).
44. Prinz, J.-H. *et al*. Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134**, 174105 (2011).
45. Noé, F. & Nuske, F. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Model. Simul.* **11**, 635–655 (2013).
46. Harrigan, M. P. *et al*. MSMBuilder: statistical models for biomolecular dynamics. *Biophys. J.* **112**, 10–15 (2017).
47. Svergun, D., Barberato, C. & Koch, M. H. CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinate. *s. J. Appl. Cryst.* **28**, 768–773 (1995).
48. Franke, D. *et al*. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Cryst.* **50**, 1212–1225 (2017).
49. Shannon, C. E. & Weaver, W. The mathematical theory of communication. *Urbana: University of Illinois Press* (1949).
50. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* **40**, 191–285 (2007).
51. Chen, P.-C. & Hub, J. S. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. *Biophys. J.* **107**, 435–447 (2014).
52. Knight, C. J. & Hub, J. S. WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicitsolvent molecular dynamics. *Nucleic Acids Res* **43**, W225–W230 (2015).

53. Pronk, S. *et al*. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
54. Kubelka, J., Chiu, T. K., Davies, D. R., Eaton, W. A. & Hofrichter, J. Sub-microsecond protein folding. *J. Mol. Biol.* **359**, 546–553 (2006).
55. Naritomi, Y. & Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *J. Chem. Phys.* **134**, 065101 (2011).
56. Nauli, S. *et al*. Crystal structures and increased stabilization of the protein g variants with switched folding pathways NuG1 and NuG2. *Protein Sci.* **11**, 2924–2931 (2002).
57. Walsh, S. T., Cheng, H., Bryson, J. W., Roder, H. & DeGrado, W. F. Solution structure and dynamics of a *de novo* designed three-helix bundle protein. *Proc. Natl. Acad. Sci. USA* **96**, 5486–5491 (1999).
58. Case, D. A. *et al. Amber 14*. University of California, San Francisco, CA (2014).
59. Berendsen, H. J., Postma, Jv, van Gunsteren, W. F., DiNola, A. & Haak, J. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
60. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).
61. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
62. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Phys. Rev. E* **87**, 012707 (2013).
63. Sheridan, R. *et al*. Evfold.org: Evolutionary couplings and protein 3d structure prediction. *Preprint at*, https://www.biorxiv.org/content/early/2015/07/02/021022 (2015).
64. Feng, J. & Shukla, D. Characterizing conformational dynamics of proteins using evolutionary couplings. *J. Phys. Chem. B* **122**, 1017–1025 (2018).
65. Rudolph, M. J., Wuebbens, M. M., Rajagopalan, K. & Schindelin, H. Crystal structure of molybdopterin synthase and its evolutionary relationship to ubiquitin activation. *Nat. Struct. Mol. Biol.* **8**, 42 (2001).
66. Hopf, T. A. *et al*. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).
67. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
68. Franke, D., Jeffries, C. M. & Svergun, D. I. Correlation map, a goodness-of-fit test for one-dimensional x-ray scattering spectra. *Nat. Methods* **12**, 419 (2015).
69. Mittal, S. & Shukla, D. Predicting optimal deer label positions to study protein conformational heterogeneity. *J. Phys. Chem. B* **121**, 9761–9770 (2017).
70. Mittal, S. & Shukla, D. Recruiting machine learning methods for molecular simulations of proteins. *Mol. Simul.* **44**, 891–904 (2018).
71. Shamsi, Z., Cheng, K. J. & Shukla, D. Reinforcement learning based adaptive sampling: REAPing rewards by exploring protein conformational landscapes. *J. Phys. Chem. B* **122**, 8386–8395 (2018).
72. Peng, J. & Zhang, Z. Unraveling low-resolution structural data of large biomolecules by constructing atomic models with experiment-targeted parallel cascade selection simulations. *Sci. Rep* **6**, 29360 (2016).
73. Weinan, E. & Vanden-Eijnden, E. Towards a theory of transition paths. *J. Stat. Phys* **123**, 503–523 (2006).
74. Weinan, E. & Vanden-Eijnden, E. Transition-path theory and path-finding algorithms for the study of rare events. *Annu. Rev. Phys. Chem.* **61**, 391–420 (2010).

Acknowledgements

We acknowledge the support from the Blue Waters sustained-petascale computing project, which is funded by the National Science Foundation (awards OCI-0725070 and ACI-1238993) and the state of Illinois. D.S. acknowledges the support from Foundation for Food and Agriculture Research via the New Innovator Award in Food & Agriculture Research. C.Z. acknowledges the support by 3M Corporate Fellowship from the Department of Chemical & Biomolecular Engineering at University of Illinois. We also thank Folding@Home for providing the HP35 folding dataset and D.E. Shaw Research for providing the protein G and the α3D folding dataset. C.Z. would like to thank Alexander S. Moffett for sharing the MoaD-MoaE association dataset. We thank Professor Jochen Hub from Saarland University for providing the modified GROMACS software package for the WAXSiS implementation. The code for implementing the adaptive sampling protocols is available at https://github.com/ShuklaGroup/SAXS-guidedAdaptiveSampling.

Author Contributions

D.S. conceived and supervised the project. C.Z. performed the simulations and analyzed the results. C.Z. and D.S. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-36090-z.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.