# *Improving Coarse-Grained Protein Force Fields with Small-Angle X-ray Scattering Data*



**Citation:** Latham, Andrew P., and Bin Zhang. "Improving Coarse-Grained Protein Force Fields with Small-Angle X-ray Scattering Data." Journal of Physical Chemistry B 123, 5 (February 2019): 957-1214 doi 10.1021/ACS.JPCB.8B10336 ©2019 Author(s)

**As Published:** 10.1021/ACS.JPCB.8B10336

**Publisher:** American Chemical Society (ACS)

**Persistent URL:** https://hdl.handle.net/1721.1/129455

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

# Improving Coarse-Grained Protein Force Fields with Small-Angle X-Ray Scattering Data

Andrew P. Latham and Bin Zhang*

*Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139*

E-mail: binz@mit.edu

Phone: 617-258-0848

## ABSTRACT

Small-angle X-ray scattering (SAXS) experiments provide valuable structural data for biomolecules in solution. We develop a highly efficient maximum entropy approach to fit SAXS data by introducing minimal biases to a coarse-grained protein force field, the associative memory, water mediated, structure and energy model (AWSEM). We demonstrate that the resulting force field, AWSEM-SAXS, succeeds in reproducing scattering profiles, and models protein structures with shapes that are in much better agreement with experimental results. Quantitative metrics further reveal a modest, but consistent improvement in the accuracy of modeled structures when SAXS data are incorporated into the force field. Additionally, when applied to a multi-conformational protein, we find that AWSEM-SAXS is able to recover the population of different protein conformations from SAXS data alone. We, therefore, conclude that the maximum entropy approach is effective in fine-tuning the force field to better characterize both protein structure and conformational fluctuation.

# INTRODUCTION

As workhorses of the cell, proteins perform a myriad of tasks via folding into various 3D conformations.[1] The importance of structures to protein function has resulted in extensive experimental efforts to their determination. Tremendous progress has been achieved in building protein structures at atomic resolution with X-ray crystallography, and many classes of protein families with different structural folds have been resolved.[2] Challenges, however, still exist for determining the structure of large proteins with multiple domains and with flexible and disordered regions.[3,4] We present an integrative approach to study protein structure and dynamics by combining computational modeling with experimental data.

Due to its efficiency and relatively low cost, computer simulation offers a promising alternative to construct structural models for proteins.[5,6] Molecular mechanics based modeling approaches are particularly powerful for studying proteins.[7–9] In addition to predicting static configurations, they can further characterize conformational fluctuation and dynamical properties that play crucial roles for protein function. Guided by the energy landscape theory,[10,11] these approaches have provided great insight into protein folding,[12–14] protein structure prediction,[15] and protein-protein interaction.[16] Central to the success of these approaches is the development of accurate force fields that describe the interaction among amino acids. Atomistic force fields derived from first principles quantum mechanics calculations are widely popular,[17–19] and recent developments in both software and hardware have significantly improved their accuracy and performance. Unfortunately, applying them for *de novo* structure prediction via direct simulations of the folding process remains impractical. Folding for most proteins occurs over $ms$ to $s$, a timescale that is beyond the accessible range of atomistic simulations. For this reason, parameterizing coarse-grained models has been a key scientific interest as they can greatly expand the accessible timescale in molecular simulations by reducing the system size.[20] In this paper, we introduce an algorithm to improve the accuracy of coarse-grained protein force fields using corrections derived from small-angle X-ray scattering (SAXS) experiments.

Integrating modeling with experiment can potentially expand the applicability of both approaches in determining high-resolution protein structures. In particular, a variety of techniques, including Cryo-electron microscopy (EM), nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), and SAXS, are becoming increasingly popular for studying systems where difficulties in sample preparation have prevented the application of X-ray diffraction. Their limited data resolution, however, renders the determination of atomic structures from these experiments alone rather challenging. Thus, numerous methods have been developed to help with their interpretation and to build structures using additional inputs provided from computational modeling.[21–24] Though significant innovation has been devoted to address the unique features of data from each type of experiment, a recurring theme in these methods is a structure-centered refinement procedure. Often, an ensemble of structures is constructed by computational modeling to select out one or multiple configurations that best fit the data. While such post-processing techniques are quite successful at deriving structures for the protein of interest, they cannot be generalized to study protein dynamics and protein-protein interactions. To overcome these limits, we adopt an energy landscape based perspective to improve the accuracy of computational models by refining their force field.

We develop an algorithm inspired by the maximum entropy principle to directly incorporate experimental constraints derived from SAXS data into a coarse-grained protein force field, associative memory, water mediated, structure and energy model (AWSEM). This algorithm is computationally efficient and provides a rigorous way to improve force field accuracy. We demonstrate that the experimentally augmented force field, which we term as AWSEM-SAXS, produces scattering profiles that are in much better agreement with experimental results than AWSEM. Furthermore, even though the improvement in the overall accuracy of modeled protein structures is modest, the simulated protein shapes as measured by the radius of gyration and extensions along the principal axes are significantly improved. Finally, we apply the algorithm to a multi-domain protein and find that it succeeds in extracting

the population of two protein conformations from the SAXS data alone. Thus incorporating experimental data into the force field can lead to more accurate energy landscapes to study both protein folding and conformational fluctuation.

# METHODS

## Maximum Entropy Approach for Improving AWSEM Force Field with SAXS

AWSEM is a coarse-grained protein force field built upon the energy landscape theory[10,11] to sculpt a funnel-like landscape with minimal frustration for well-folded proteins. It utilizes a combination of physically motivated and bioinformatics based potentials to recreate protein structure with only three sites per amino acid.[14,15,25–28] AWSEM has been successfully applied to study protein-protein interaction,[16] protein aggregation,[29] and membrane protein folding;[30,31] meanwhile, it has been extended for protein-DNA complexes[32–34] and intrinsically disordered proteins.[35] Our study further improves the accuracy of the AWSEM force field with SAXS data through the maximum entropy principle.[36]

SAXS is an effective way of obtaining low-resolution structural data for biomolecules and offers a promising alternative for characterizing protein structures in solution.[37,38] The resolution of the SAXS technique is limited by an orientational average, and its output is often represented as a one-dimensional intensity profile. These profiles provide valuable information about the shapes and dimensions of protein molecules. Through an indirect Fourier transform, the scattering profile is related to the pair-wise distance distribution function of electrons, $p(r)$, which can be viewed as a series of ensemble averages

$$p(r_k)\Delta r = \langle \mathcal{N}(r_k) \rangle_{\mathrm{SAXS}} = \left\langle \frac{2}{N(N-1)} \sum_{i>j} \Theta_k(r_{ij}) \right\rangle, \quad \text{for } k = 1, \cdots, n, \tag{1}$$

where $r_{ij}$ is the distance between a pair of electrons $i$ and $j$, $N$ refers to the total number

of electrons, $\Delta r \approx 1.5$Å measures the size of the bin used for calculating the distance distribution, and $n$ is the total number of bins. The switching function

$$\Theta_k(r_{ij}) = \frac{1}{4}\{1 + \tanh[\eta(r_{ij} - r_{\min})]\}\{1 + \tanh[\eta(r_{\max} - r_{ij})]\}, \tag{2}$$

where $r_{\min} = r_k - \frac{\Delta r}{2}$, $r_{\max} = r_k + \frac{\Delta r}{2}$, and $\eta = 7\text{Å}^{-1}$. $\Theta_k(r_{ij})$ is equal to 1 in a narrow region around $r_k$ and 0 otherwise. The angular brackets in Eq. 1 represents an average over all the conformations adopted by the molecule in solution. A corresponding quantity as that in Eq. 1 can be defined in computer simulations, $\langle \mathcal{N}(r_k) \rangle_{\text{sim}}$, by calculating the average over the simulated conformational ensemble. For coarse-grained models, an approximation for $\langle \mathcal{N}(r_k) \rangle_{\text{sim}}$, which was shown to be sufficient to reproduce SAXS data in previous studies,[39–41] can be obtained using pair-wise distances between $\alpha$-carbons only instead of electrons. If the simulation accurately models the structure(s) for the molecule of interest, then $\langle \mathcal{N}(r_k) \rangle_{\text{sim}} \equiv \langle \mathcal{N}(r_k) \rangle_{\text{SAXS}} \equiv p(r_k)\Delta r$, for $k = 1, \cdots, n$; otherwise, the discrepancy between simulation and experiment can be used to improve the accuracy of the force field used in simulation.

To improve the agreement between simulated and experimental scattering profiles and the accuracy of structures modeled by AWSEM, we introduce biases derived from the maximum entropy principle to the force field. Maximum entropy approaches provide the minimally biased choice of restraint potential to nudge simulations toward experimental measurements,[36,42–44] and have been successfully applied to study a wide range of problems.[45–48] Beginning with the AWSEM force field, the overall energy that maximizes the information entropy while reproducing the experimental scattering profile is

$$\mathcal{H}_{\text{AWSEM-SAXS}} = \mathcal{H}_{\text{AWSEM}} + \sum_{k=1}^{n} \alpha_k \mathcal{N}(r_k) \tag{3}$$

where $\{\alpha_k\}$ are external parameters whose values can be fine tuned to ensure that $\langle \mathcal{N}(r_k) \rangle_{\text{sim}} \equiv \langle \mathcal{N}(r_k) \rangle_{\text{SAXS}}$ for $k = 1, \cdots, n$.

It has been shown previously,[36,45] that the values for the list of $\boldsymbol{\alpha} = \{\alpha_k\}$ can be found

by taking extrema of the approximate objective function

$$\Gamma(\boldsymbol{\alpha}) = \frac{\beta^2}{2}\boldsymbol{\alpha}^T * \boldsymbol{B} * \boldsymbol{\alpha} - \beta\left[\langle\mathcal{N}(\boldsymbol{r})\rangle_{\text{sim}} - \langle\mathcal{N}(\boldsymbol{r})\rangle_{\text{SAXS}}\right]^T\boldsymbol{\alpha} \tag{4}$$

where $T$ represents the transpose operator, and $\beta$ is the inverse temperature. $\boldsymbol{B}$ is the covariance matrix with the elements $B_{kl} = \langle\mathcal{N}(r_k)\mathcal{N}(r_l)\rangle_{\text{sim}} - \langle\mathcal{N}(r_k)\rangle_{\text{sim}}\langle\mathcal{N}(r_l)\rangle_{\text{sim}}$.

Following Zhang and Wolynes,[45] we used an iterative expression to update $\boldsymbol{\alpha}^{t+1}$ based on values from a previous iteration $\boldsymbol{\alpha}^t$ and ensemble averages calculated with those parameters

$$\boldsymbol{\alpha}^{t+1} = \boldsymbol{\alpha}^t + \frac{1}{\beta}\boldsymbol{B}^{-1}[\langle\mathcal{N}(\boldsymbol{r})\rangle_{\text{sim}} - \langle\mathcal{N}(\boldsymbol{r})\rangle_{\text{SAXS}}]^T. \tag{5}$$

Specifically, we used the following three steps to determine the maximum entropy biases. 1) Perform simulations with $\mathcal{H}_{\text{AWSEM-SAXS}}$ to collect an ensemble of protein structures and estimate the averages $\langle\mathcal{N}(\boldsymbol{r})\rangle_{\text{sim}}$ and $\boldsymbol{B}$. 2) Calculate the relative error between simulated and experimental pair-wise distance distributions defined as

$$\epsilon = \frac{\sum_k |\langle\mathcal{N}(r_k)\rangle_{\text{sim}} - \langle\mathcal{N}(r_k)\rangle_{\text{SAXS}}|}{\sum_k \langle\mathcal{N}(r_k)\rangle_{\text{SAXS}}} \tag{6}$$

to see if it is lower than a user-defined tolerance. 3) If $\epsilon$ is less than the tolerance, the force field has been successfully updated to incorporate SAXS data. If not, $\boldsymbol{\alpha}$ will be updated according to Eq. 5 to carry out another round of iteration. The entire protocol is summarized as a flow chart in Figure 1.

## Simulation Details

To calculate the ensemble averages, we carried out constant temperature and volume simulations (NVT) using the molecular dynamics package Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS).[49] Simulations were initialized with extended protein configurations and lasted for 1500000 steps, with the first 500000 steps discarded for equilibration.

Initial protein configurations were obtained from a high temperature simulation conducted at 1000K. Note that the time scale in coarse grained models is typically not well defined, and one simulation step in similar models has been mapped to 1 *ps* by matching the diffusion coefficient of protein domains between coarse grained and all atom simulations.[50] We further applied the parallel tempering technique[51] and utilized four independent tempering trajectories to enhance the efficiency of conformational sampling. A total of six replicas with equally spaced temperatures from 300K to 500K were used for each tempering trajectory.

Default parameters in the AWSEM force field, which can be found in Table S1, were used to conduct the simulations unless otherwise specified. A key component of the force field is a bioinformatics inspired memory term that biases consecutive fragments of nine residues in length toward a set of conformations with high sequence similarity collected from the Protein Data Bank (PDB). Homologues were excluded when building this dataset to mirror a *de novo* prediction.

To evaluate the accuracy of force fields in modeling protein structures, we performed simulated annealing calculations. Starting again from extended protein configurations produced from the high temperature simulation mentioned above, Langevin dynamics reduced the temperature from 500K to 200K over 8000000 steps. A total of 20 runs were performed for each protein to ensure statistical significance.

## SAXS Data

To provide a comprehensive evaluation of the performance of AWSEM-SAXS, we applied it to a diverse set of proteins. This set covers a large range of amino acid lengths, and includes proteins whose structures AWSEM is known to model both well and poorly. Details for all the proteins are summarized in Table S2. Unfortunately, not all of these proteins have not been studied with SAXS. For proteins without experimental data, we used the following protocol to produce simulated SAXS profiles. First, we performed an NVT simulation of 1000000 steps, during which the protein is restrained to the PDB structure using an amh-Gō potential[52]

in addition to the standard AWSEM force field. From the simulated structural ensemble in which the protein closely resembles the crystal structure, we determined the average pair-wise distance distribution $\langle \mathcal{N}(\boldsymbol{r}) \rangle_{\mathrm{SAXS}}$ using $\alpha$-carbons only. This approximation maintains residue-level accuracy, greatly reduces computational time, and has provided good results in previous studies.[39–41] For proteins with experimental data, we downloaded their scattering profiles and pair-wise distance distributions from the small-angle scattering biological data bank (SASBDB).[53] We also ensured that high resolution structures are available for all proteins to evaluate the accuracy of modeled protein conformations using metrics defined below.

## Structural Analysis

We used four metrics, fraction of native contacts ($Q$), root mean squared displacement (RMSD), radius of gyration ($R_{\mathrm{g}}$), and principal moments of initia or extensions along principal axes ($\boldsymbol{\lambda}$), to evaluate the accuracy of modeled structures. Coordinates of $\alpha$-carbons were used to calculate these quantities. $Q$ measures the similarity between a modeled structure with the PDB conformation and takes values between 0 and 1, with larger values corresponding to agreement. It takes the form

$$Q = \frac{2}{(N-2)(N-3)} \sum_{i<j-2} \exp\left( - \frac{(r_{ij} - r_{ij}^{o})^2}{2\sigma_{ij}^2} \right), \tag{7}$$

where $r_{ij}^{o}$ is the pair-wise distance between $\alpha$-carbons $i$ and $j$ in the crystal structure, $N$ is the number of amino acids, $i$ and $j$ are amino acids separated by a distance $r_{ij}$, and $\sigma_{ij} = (1 + |i-j|)^{0.15}$.

RMSD similarly measures the difference between a modeled structure and a reference structure. After removing translations and rotations,

$$\mathrm{RMSD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} ||\vec{r_i} - \vec{r_i^{o}}||^2}, \tag{8}$$

where $\vec{r_i}$ and $\vec{r_i^o}$ are the positions of the $i$-th $\alpha$-carbon in the current and the reference structure and $N$ is the number of amino acids. The double bars denote the Euclidean norm.

$R_{\rm g}$ is a measurement of protein size, and takes the form

$$R_{\rm g} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} ||\vec{r_i} - \vec{r}_{\rm com}||^2}, \tag{9}$$

where $\vec{r}_{\rm com}$ is the average position of all $\alpha$-carbons and other symbols are similarly defined as in Eq. 8.

$\boldsymbol{\lambda} = \{\lambda_x, \lambda_y, \lambda_z\}$ measures the extension of a protein along its principal axes and provides a more comprehensive characterization of protein shape than $R_{\rm g}$. $\boldsymbol{\lambda}$ can be determined as eigenvalues of the inertia matrix $\mathbf{I}$, whose elements are defined as

$$I_{kl} = \frac{1}{N}\sum_{i=1}^{N}(r_i^k - r_{\rm com}^k)(r_i^l - r_{\rm com}^l), \tag{10}$$

where $k, l = 1, 2, 3$ are indices of the three Cartesian axis, $r_{\rm com}$ is the protein center of mass, and $i$ loops over all the $\alpha$-carbons. Denoting $\lambda_z$ as the one that is along the most deviant axis and differs the most from the mean of the three eigenvalues, we then define the shape anisotropy ($\Delta$) as

$$\Delta = |\lambda_z - \frac{1}{2}(\lambda_x + \lambda_y)|. \tag{11}$$

We further normalize $\Delta$ by $R_{\rm g}^2$ to remove protein size dependence.

Additionally, we computed SAXS curves for proteins with experimental scattering profiles for a direct comparison. A total of 4000 representative structures were selected from the simulated conformational ensemble at 300K to calculate SAXS profiles using CRYSOL.[54]

# RESULTS AND DISCUSSION

## Optimization Algorithm Succeeds in Reproducing SAXS Data

As a proof of principle, we first applied the maximum entropy approach on a small protein with two helices (PDB ID: 4djg). The AWSEM force field alone succeeds in modeling the structure of this protein. For testing purposes, we therefore turned off tertiary interactions in the force field, the water-mediated potential. In addition, we only included fragment memory terms for the helical, but not the hinge, regions of the protein. The resulting potential $\mathcal{H}_{\text{initial}}$ poorly describes the protein structure, and the simulated $\langle \mathcal{N}(\boldsymbol{r}) \rangle_{\text{sim}}$ differs significantly from the "experimental" one created using our protocol detailed in the *Methods* section.

Starting from $\mathcal{H}_{\text{initial}}$, we performed the optimization procedure to derive the biasing terms to the force field. As shown in Figure 2A, the difference between simulated and experimental probability distribution for pair-wise distances decreases quickly, and the error defined in Eq. 6, $\epsilon$, drops to less than 6% during our optimization procedure. The best structure modeled by the optimized Hamiltonian using annealing simulations is in good agreement with the PDB structure with a RMSD less than 1 Å.

## Augmenting AWSEM with SAXS Data Improves Structure Modeling

To more systematically evaluate the performance of AWSEM-SAXS in structure modeling, we applied it to eleven proteins that fell into two groups. The first group consists of proteins studied in previous AWSEM studies.[15,55] Since these proteins have not been studied using SAXS, we simulated pair-wise distance distributions according to the procedure outlined in the *Methods* section. For the other set of proteins that consists of 1soy and 3mzq, both PDB and SAXS data are available. We carried out the iterative optimization algorithm to derive the force field correction terms for each protein. As shown in Figure S1, we were able to reach an error $\epsilon$ between simulated and experimental pair-wise distance distribution (see Eq.

6) less than 5% for the first group, and 10% for the second group.

For a more direct comparison, we further constructed the scattering profiles using conformations collected at 300K for the two proteins with experimental SAXS data. Figures 3A and 3B demonstrate that, for both proteins, AWSEM-SAXS produces profiles that are in better agreement with experimental ones than those obtained with AWSEM, especially in the small $s$ (large distance) regime. We further quantified the similarity between simulated and experimental scattering profiles using $\chi^2$ defined below

$$\chi^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{I_{\text{exp}}(s_i) - cI_{\text{sim}}(s_i)}{\sigma(s_i)}\right)^2 \tag{12}$$

where $N$ is the total number of data points. $I_{\text{exp}}$ and $I_{\text{sim}}$ are the experimental and simulated scattering profiles, and $\sigma$ is the standard deviation of the experimental profile. $c$ is a scaling factor tuned to minimize $\chi^2$. We computed the overall $\chi^2$ as well as $\chi^2_{0.15}$ that only includes data for the region where $s < 0.15$. The results confirm our qualitative observations and are summarized in Table 1.

To examine the overall quality of protein configurations modeled by AWSEM-SAXS versus AWSEM, we determined the mean and standard deviation of the radius of gyration, $R_{\text{g}}$, using the ensemble of structures collected at 300 K. We further calculated experimental $R_{\text{g}}$ directly from SAXS data when available, and from the PDB structures otherwise. As shown in Figure 3C, AWSEM (green) generally over-predicts the size of the studied proteins, as indicated by large deviations from the experimental values (blue). On the other hand, with SAXS biasing (red), we consistently reproduce experimental $R_{\text{g}}$ with high accuracy, and therefore provide a much better description for protein size. We further characterized the shapes of modeled protein conformations by analyzing the extensions along their principal axes and calculating the anisotropy measure $\Delta/R_{\text{g}}^2$. As shown in Figure 3D and S3A, compared with AWSEM-SAXS, AWSEM over estimates the size along the longest axis. Consistent with these observations, the Pearson correlation coefficient between $\Delta/R_{\text{g}}^2$ calculated

using PDB and AWSEM-SAXS structures (0.85) is significantly higher than that between PDB and AWSEM structures (0.57) (see S3B-D). It is important to note that these improvements in protein shape characterization hold true for proteins optimized using both simulated and experimental SAXS data.

Finally, we carried out simulated annealing simulations for each protein using the two force fields to compare the performance of top-ranking structures. The highest $Q$ and lowest RMSD values along each simulation are shown in Figure 4, with red and blue for AWSEM-SAXS and AWSEM results respectively. For most proteins, AWSEM-SAXS outperforms in providing structures with larger $Q$ and smaller RMSD. These results suggest that incorporating SAXS data into the force field leads to a modest, but consistent improvement in the accuracy of modeled structures. We further tracked protein configurations with lowest energies along these simulations. As shown in Figure S2, we again observe a systematic improvement in the similarity between simulated and PDB structures using AWSEM-SAXS.

Figure 4 also shows that the improvement in structural modeling is rather heterogeneous. For example, while AWSEM-SAXS significantly outperforms AWSEM for protein 256b, the two force fields are comparable for protein 1n2x. To better understand the uneven performance for different proteins, we computed the contact maps for modeled structures. As shown in Figure 5A, protein 256b adopts a rather simple fold as a four helix bundle. AWSEM alone succeeds in predicting the secondary structures (left panel). The improved compactness from incorporating SAXS data translates well into a better model for the tertiary structure. On the other hand, protein 1n2x has a more complex fold with long loop regions that are not well predicted by AWSEM. The poor quality of secondary structures is not rectified by AWSEM-SAXS due to the low resolution of SAXS data. Therefore, the accuracy of the overall structure does not see a dramatic change despite a significant improvement in protein shape. The slight decrease in $Q$ for protein 1n2x is because that Figure 4 reports the highest value over all the conformations explored along the simulated annealing trajectories. AWSEM-SAXS force field is less flexible due to the added constraints, and

therefore has more limited conformational exploration compared to AWSEM. These results suggest that better secondary structure predictions could potential result in more significant improvement in modeling protein tertiary structures using AWSEM-SAXS.

## Incorporating SAXS Data Enables Fine-Tuning of the Energy Landscape

The results presented so far suggest that the maximum entropy approach, when applied to proteins with unique conformations, can improve the accuracy of modeled structures. Here, we further apply this approach to a multi-conformational protein and demonstrate that the relative population of different conformations can be decoded from SAXS data as well. SAXS profiles are ensemble averages determined using all possible protein conformations, and therefore, encode the population information. Previous structure-centered post-processing approaches were indeed able to extract this information for different clusters of protein conformations.[23,56] Quantifying the conformational population is of significant interest since it can help distinguish the two different mechanisms of protein function, i.e., conformational selection versus induced fit.[57]

As a proof of principle, we engineered a bistable system using the same protein studied in Figure 2. In its native state, which we will refer to as Structure 1, the two helices of protein 4djg are in contact, as can be seen in Figure 6A. As mentioned previously, the AWSEM force field stabilizes the native state, and only exhibits a single basin. To create a second energy basin, we introduced a Gaussian bias to the force field,

$$\mathcal{H}_{\text{Gaussian}} = -k\, e^{-(\text{RMSD}_2-2)^2/4}, \tag{13}$$

where $k = 20$ kcal/mol. $\text{RMSD}_2$ is the minimal root mean squared displacement from Structure 2, in which the alpha helices extend in opposite directions, as shown in the right panel of Figure 6A. Our initial Hamiltonian becomes $\mathcal{H}_{\text{initial}} = \mathcal{H}_{\text{AWSEM}} + \mathcal{H}_{\text{Gaussian}}$.

Starting from $\mathcal{H}_{\text{initial}}$, we performed a series of optimizations in which the force field is steered toward simulated SAXS data calculated using different populations of the two structures shown in Figure 6A. These "experimental" distributions were created using linear combinations of the pair-wise distance distributions calculated from simulations constrained to each protein structure

$$\langle \mathcal{N}(\boldsymbol{r}) \rangle_{\text{SAXS}} = p_1 \langle \mathcal{N}(\boldsymbol{r}) \rangle_1 + p_2 \langle \mathcal{N}(\boldsymbol{r}) \rangle_2 . \tag{14}$$

We studied a total of three states where $p_1 = 0, 0.5$ and 1.0 and $p_2 = 1 - p_1$ respectively. The resulting probability distributions are shown in Figure 6B and are compared to those obtained from optimized force fields in Figure S4.

To monitor the ensemble of modeled protein conformations, we measured $Q$ and RMSD with respect to each structure, as denoted by the subscript. In addition, we tracked an angle $\theta$ between the 4th, 21st, and 40th amino acids. As illustrated in Figure 6A, the angle easily distinguishes the two conformations with values of $24^o$ and $172^o$. Figures 6C, 6D and 6E present the probability distribution for $\theta$, $Q_2 - Q_1$ and $\text{RMSD}_2 - \text{RMSD}_1$ calculated using AWSEM-SAXS derived for different mixtures of protein conformations (see Figure S5 for individual distributions of $\text{RMSD}_2, \text{RMSD}_1, Q_2$, and $Q_1$). From these distributions, it is clear that the resulting force fields succeed in capturing the dramatic shift in population of different protein conformations. Assuming that protein conformations with $\text{RMSD}_2 - \text{RMSD}_1$ less than zero correspond to the Structure 1 ensemble, we obtained populations of 6.0%, 53.6% and 100% for $p_1 = 0$, $p_1 = 0.5$, and $p_1 = 1$ respectively. Therefore, we were able to quantitatively recover the population information encoded in the SAXS data within 6%. Using $\theta$ or $Q$ instead of RMSD to measure the ensemble population results in similar conclusions. These results demonstrate our method's potential for quantitatively adjusting the energy landscape for modeling multi-domain proteins.

# CONCLUSIONS

This study utilized the maximum entropy principle to bias coarse grained MD simulations toward experimental SAXS profiles. We demonstrated that our biased force field, AWSEM-SAXS, models protein shapes that are in much better agreement with experimental structures. Furthermore, it improves structure modeling by finding a higher $Q$ and lower RMSD for most proteins studied when compared with the original AWSEM force field. In addition to well-folded proteins, we applied the maximum entropy approach to a multi-conformational protein, and found that incorporating SAXS data into the force field leads to a fine tuning of the energy landscape to match the conformational population. In the studied case, we were able to recreate a mixture of extended and closed protein conformations within 6% of the real distribution. Since protein conformational dynamics are dictated by the underlying energy landscape, we anticipate that AWSEM-SAXS may lead to more accurate modeling of dynamical quantities as well. To rigorously address the shortcoming of simulating kinetic properties in general, however, might require explicit incorporation of dynamical measurements into the model.[58]

Our results further suggest that incorporating SAXS data into the force field will be most useful when protein secondary structures are known or can be well predicted by the force field. SAXS data itself cannot provide such information due to the low resolution. Fortunately, in many cases, including intrinsically disordered proteins, multi-domain proteins, and protein protein complexes, the structures for individual monomers or domains can be or have been determined. For these systems, we anticipate the maximum entropy approach developed here will find its most success by correctly modeling the shape of the overall complex. As experimental biases are directly incorporated into the force field, AWSEM-SAXS can be applied to study the aggregation of intrinsically disordered proteins and to probe the sequence features of these proteins in driving the formation of phase-separated liquid droplets.

Finally, we emphasize that the maximum entropy approach is not restricted to the AWSEM potential or even to proteins. Implemented in the LAMMPS collective variable

module,[59] the maximum entropy correction terms defined in Eq. 3 may be combined with other force fields for dynamical simulation of any bimolecular molecule where SAXS data is available.

# SUPPORTING INFORMATION DESCRIPTION

The Supporting Information includes additional simulation details and analysis and is available free of charge.

# ACKNOWLEDGEMENT

This work was supported by startup funds from the Department of Chemistry at the Massachusetts Institute of Technology. A.L. acknowledges the Jan S. (1973) and Ruby Krouwer Fellowship for financial support.

# References


(1) Lodish, H.; Berk, A.; Kaiser, C. A.; Matsudaira, P.; Krieger, M.; Scott, M. P.; Darnell, J.; Others, *Molecular Cell Biology*; W. H. Freeman, 2004.

(2) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(3) Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.

(4) Han, J.-H.; Batey, S.; Nickson, A. A.; Teichmann, S. A.; Clarke, J. The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 319.

(5) Bonneau, R.; Baker, D. Ab Initio Protein Structure Prediction: Progress and Prospects. *Annu. Rev. Biophys. Biomol. Struct.* **2001**, *30*, 173–189.

(6) Petrey, D.; Honig, B. Protein structure prediction: inroads to biology. *Mol. Cell* **2005**, *20*, 811–819.

(7) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646.

(8) Adcock, S. A.; McCammon, J. A. Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins. *Chem. Rev.* **2006**, *106*, 1589–1615.

(9) Dror, R. O.; Dirks, R. M.; Grossman, J. P.; Xu, H.; Shaw, D. E. Biomolecular Simulation: a Computational Microscope for Molecular Biology. *Annu. Rev. Biophys.* **2012**, *41*, 429–452.

(10) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **1995**, *21*, 167–195.

(11) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of Protein Folding: the Energy Landscape Perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545–600.

(12) Onuchic, J. N.; Wolynes, P. G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75.

(13) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.

(14) Schafer, N. P.; Kim, B. L.; Zheng, W.; Wolynes, P. G. Learning to fold proteins using energy landscape theory. *Isr. J. Chem.* **2014**, *54*, 1311–1337.

(15) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **2012**, *116*, 8494–8503.

(16) Zheng, W.; Schafer, N. P.; Davtyan, A.; Papoian, G. A.; Wolynes, P. G. Predictive energy landscapes for protein-protein association. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 19244–19249.

(17) MacKerell, A. D. et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(18) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(19) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(20) Saunders, M. G.; Voth, G. A. Coarse-Graining Methods for Computational Biology. *Annu. Rev. Biophys.* **2013**, *42*, 73–93.

(21) Trabuco, L. G.; Villa, E.; Mitra, K.; Frank, J.; Schulten, K. Flexible Fitting of Atomic Structures into Electron Microscopy Maps Using Molecular Dynamics. *Structure* **2008**, *16*, 673–683.

(22) Różycki, B.; Kim, Y. C.; Hummer, G. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* **2011**, *19*, 109–116.

(23) Boura, E.; Rozycki, B.; Herrick, D. Z.; Chung, H. S.; Vecer, J.; Eaton, W. A.; Cafiso, D. S.; Hummer, G.; Hurley, J. H. Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 9437–9442.

(24) Robustelli, P.; Kohlhoff, K.; Cavalli, A.; Vendruscolo, M. Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. *Structure* **2010**, *18*, 923–933.

(25) Hardin, C.; Eastwood, M. P.; Prentiss, M. C.; Luthey-Schulten, Z.; Wolynes, P. G. Associative memory Hamiltonians for structure prediction without homology: $\alpha/\beta$ proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 1679–1684.

(26) Papoian, G. A.; Ulander, J.; Eastwood, M. P.; Luthey-Schulten, Z.; Wolynes, P. G. Water in protein structure prediction. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 3352–7.

(27) Friedrichs, M.; Wolynes, P. G. Toward Protein Tertiary Structure Recognition by Means of Associative Memory Hamiltonians. *Science* **1989**, *246*, 371–3.

(28) Papoian, G. *Coarse-Grained Modeling of Biomolecules*; Series in Computational Biophysics; CRC Press, 2017; pp 1–60.

(29) Zheng, W.; Schafer, N. P.; Wolynes, P. G. Free energy landscapes for initiation and branching of protein aggregation. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 20515–20520.

(30) Truong, H. H.; Kim, B. L.; Schafer, N. P.; Wolynes, P. G. Predictive energy landscapes for folding membrane protein assemblies. *J. Chem. Phys.* **2015**, *143*, 243101.

(31) Kim, B. L.; Schafer, N. P.; Wolynes, P. G. Predictive energy landscapes for folding alpha-helical transmembrane proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2014**, *111*, 11031–11036.

(32) Tsai, M. Y.; Zhang, B.; Zheng, W.; Wolynes, P. G. Molecular Mechanism of Facilitated Dissociation of Fis Protein from DNA. *J. Am. Chem. Soc.* **2016**, *138*, 13497–13500.

(33) Zhang, B.; Zheng, W.; Papoian, G. A.; Wolynes, P. G. Exploring the Free Energy Landscape of Nucleosomes. *J. Am. Chem. Soc.* **2016**, *138*, 8126–8133.

(34) Potoyan, D. A.; Zheng, W.; Komives, E. A.; Wolynes, P. G. Molecular stripping in the NF-$\kappa$B/I$\kappa$B/DNA genetic regulatory network. *Proc. Natl. Acad. Sci.* **2016**, *113*, 110–115.

(35) Wu, H.; Zhao, H.; Wolynes, P. G.; Papoian, G. A. AWSEM-IDP : A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J. Phys. Chem. B* **in press**,

(36) Pitera, J. W.; Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445–3451.

(37) 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.* **2007**, *40*, 191–285.

(38) Koch, M. H. J.; Vachette, P.; Svergun, D. I. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q. Rev. Biophys.* **2003**, *36*, 147–227.

(39) Zheng, W.; Doniach, S. Protein structure prediction constrained by solution X-ray scattering data and structural homology identification. *J. Mol. Biol.* **2002**, *316*, 173–187.

(40) Wu, Y.; Tian, X.; Lu, M.; Chen, M.; Wang, Q.; Ma, J. Folding of small helical proteins assisted by small-angle X-ray scattering profiles. *Structure* **2005**, *13*, 1587–1597.

(41) Yang, S.; Park, S.; Makowski, L.; Roux, B. A rapid coarse residue-based computational method for X-ray solution scattering characterization of protein folds and multiple conformational states of large protein complexes. *Biophys. J.* **2009**, *96*, 4449–4463.

(42) Roux, B.; Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **2013**, *138*, 084107.

(43) Cesari, A.; Reißer, S.; Bussi, G. Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments. *Computation* **2018**, *6*, 1–26.

(44) Dannenhoffer-Lafage, T.; White, A. D.; Voth, G. A. A Direct Method for Incorporating Experimental Data into Multiscale Coarse-Grained Models. *J. Chem. Theory Comput.* **2016**, *12*, 2144–2153.

(45) Zhang, B.; Wolynes, P. G. Topology, structures, and energy landscapes of human chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 6062–6067.

(46) Zhang, B.; Wolynes, P. G. Shape Transitions and Chiral Symmetry Breaking in the Energy Landscape of the Mitotic Chromosome. *Phys. Rev. Lett.* **2016**, *116*, 248101.

(47) Zhang, B.; Wolynes, P. G. Genomic Energy Landscapes. *Biophys. J.* **2017**, *112*, 1–7.

(48) Di Pierro, M.; Zhang, B.; Aiden, E. L.; Wolynes, P. G.; Onuchic, J. N. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 12168–12173.

(49) Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(50) Takada, S.; Kanada, R.; Tan, C.; Terakawa, T.; Li, W.; Kenzaki, H. Modeling Structural Dynamics of Biomolecular Complexes by Coarse-Grained Molecular Simulations. *Acc. Chem. Res.* **2015**, *48*, 3026–3035.

(51) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(52) Eastwood, M. P.; Wolynes, P. G. Role of explicitly cooperative interactions in protein folding funnels: a simulation study. *J. Chem. Phys.* **2001**, *114*, 4702–4716.

(53) Valentini, E.; Kikhney, A. G.; Previtali, G.; Jeffries, C. M.; Svergun, D. I. SASBDB, a repository for biological small-angle scattering data. *Nucleic Acids Res.* **2015**, *43*, D357–D363.

(54) Svergun, D.; Barberato, C.; Koch, M. H. CRYSOL - A program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* **1995**, *28*, 768–773.

(55) Sirovetz, B. J.; Schafer, N. P.; Wolynes, P. G. Protein structure prediction: making AWSEM AWSEM-ER by adding evolutionary restraints. *Proteins: Struct., Funct., Bioinf.* **2017**, *85*, 2127–2142.

(56) Schneidman-Duhovny, D.; Hammel, M.; Tainer, J. A.; Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **2013**, *105*, 962–974.

(57) Boehr, D. D.; Nussinov, R.; Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **2009**, *5*, 789.

(58) Chen, J.; Chen, J.; Pinamonti, G.; Clementi, C. Learning Effective Molecular Models from Experimental Observables. *J. Chem. Theory Comput.* **2018**, *14*, 3849–3858.

(59) Fiorin, G.; Klein, M. L.; Hénin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **2013**, *111*, 3345–3362.

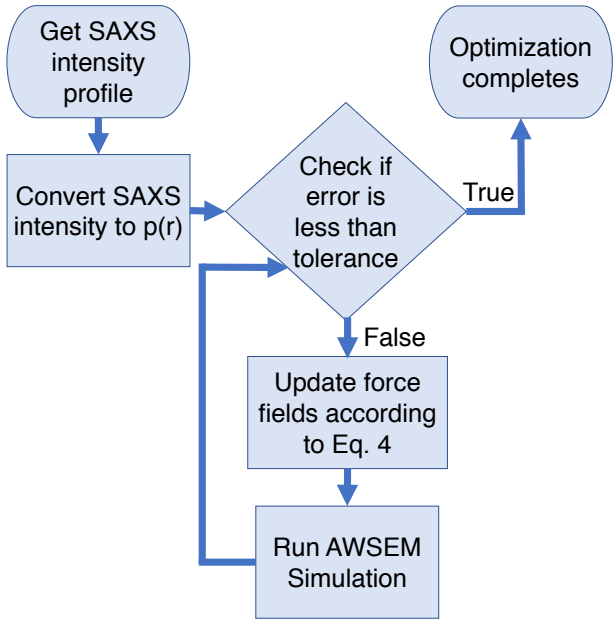


Figure 1: Illustration of the iterative optimization algorithm to bias the AWSEM force field with SAXS data.

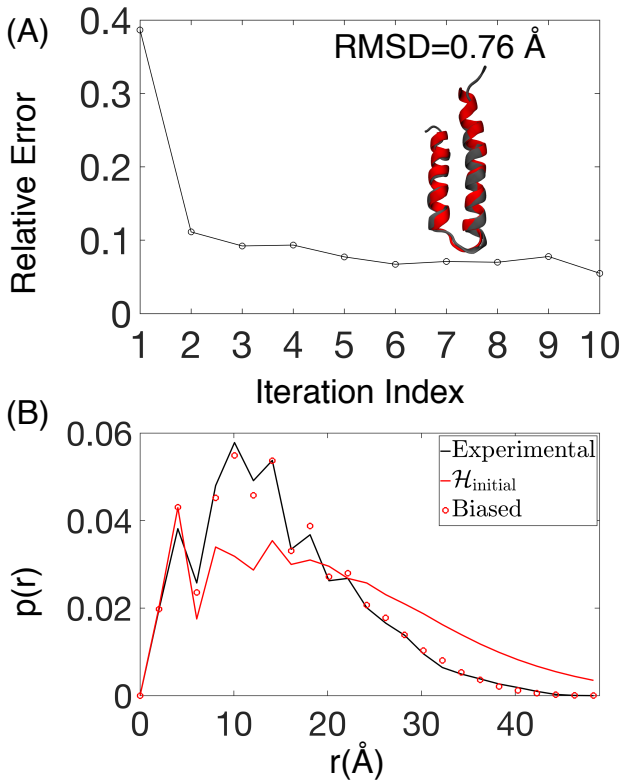


Figure 2: The maximum entropy approach provides a force field that succeeds in structure modeling and reproducing SAXS-data. (A) The relative error between experimental and simulated pair-wise distance distribution as a function of iterations of the optimization algorithm. The insert shows the similarity between the modeled (red) and the PDB (gray) structure. (B) Comparison between simulated pair-wise distance distributions calculated using the initial (red line) and biased force field (red dots) with the experimental one (black).

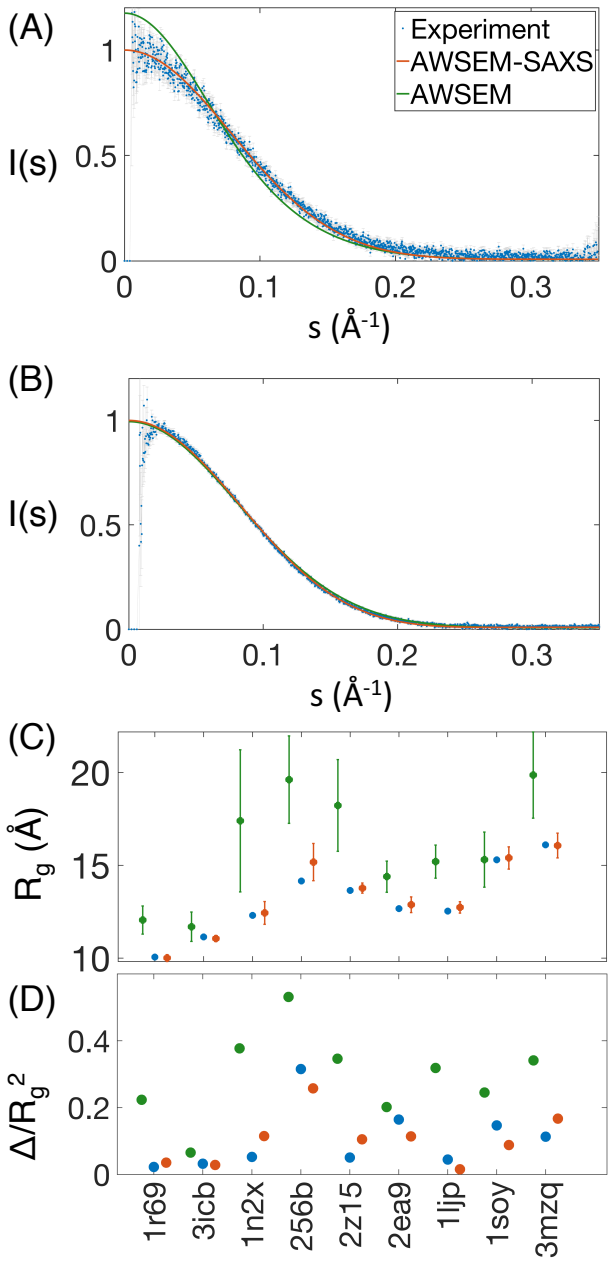


Figure 3: Incorporating experimental biases into the force field improves the modeling of scattering profiles and protein shapes. (A) Comparison between scattering profiles for protein 3mzq calculated with structure models from AWSEM (green) and from AWSEM-SAXS (red). The experimental profile is shown in blue dots, with the standard deviations shown in grey. (B) The same data as in part (A) but for protein 1soy. (C) Comparison of the radius of gyration for experimental (blue), AWSEM (green) and AWSEM-SAXS (red) protein structures. The error bars correspond to standard deviations from the mean. (D) Comparison of the shape anisotropy for different protein structures. The coloring scheme is the same as in part (C).

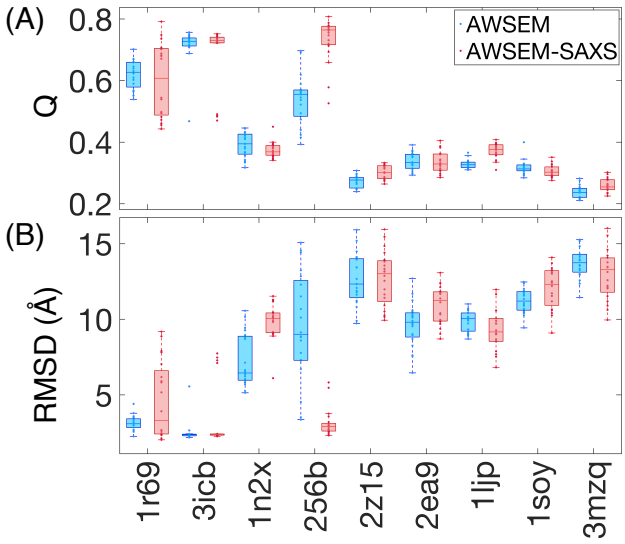


Figure 4: The highest $Q$ (A) and lowest RMSD (B) values for protein structures explored in 20 simulated annealing simulations carried out with the AWSEM (blue) and AWSEM-SAXS (red) force field. The boxes represent the 25% and 75% quantities of the distribution, and the line inside each box corresponds to the median value. Whiskers indicate the last values that fall within 1.5 times the interquartile range.

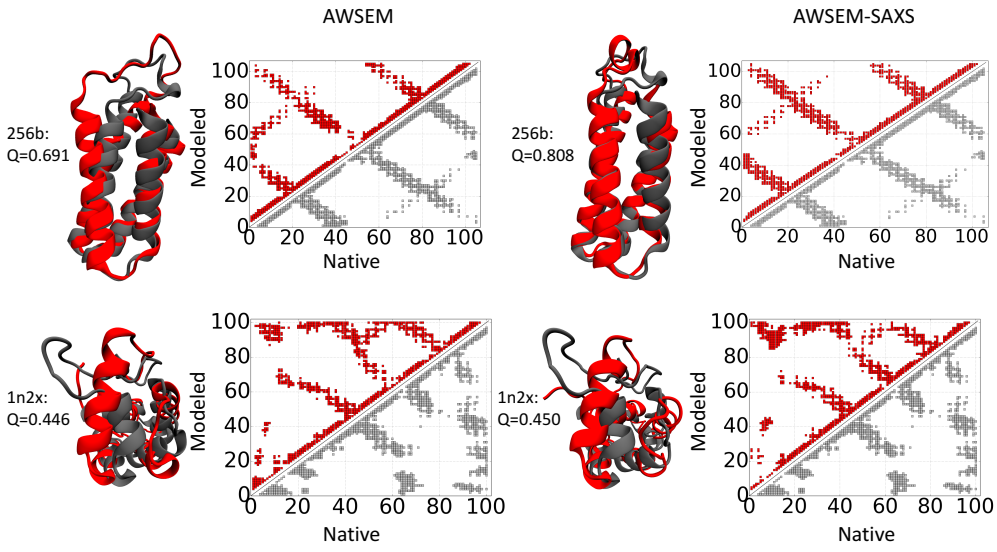


Figure 5: Comparison between contact maps calculated using highest-$Q$-value AWSEM structures (left panels) and AWSEM-SAXS structures (right panels) with that for PDB structures (grey). 3D rendering of the corresponding structures are shown on the side.

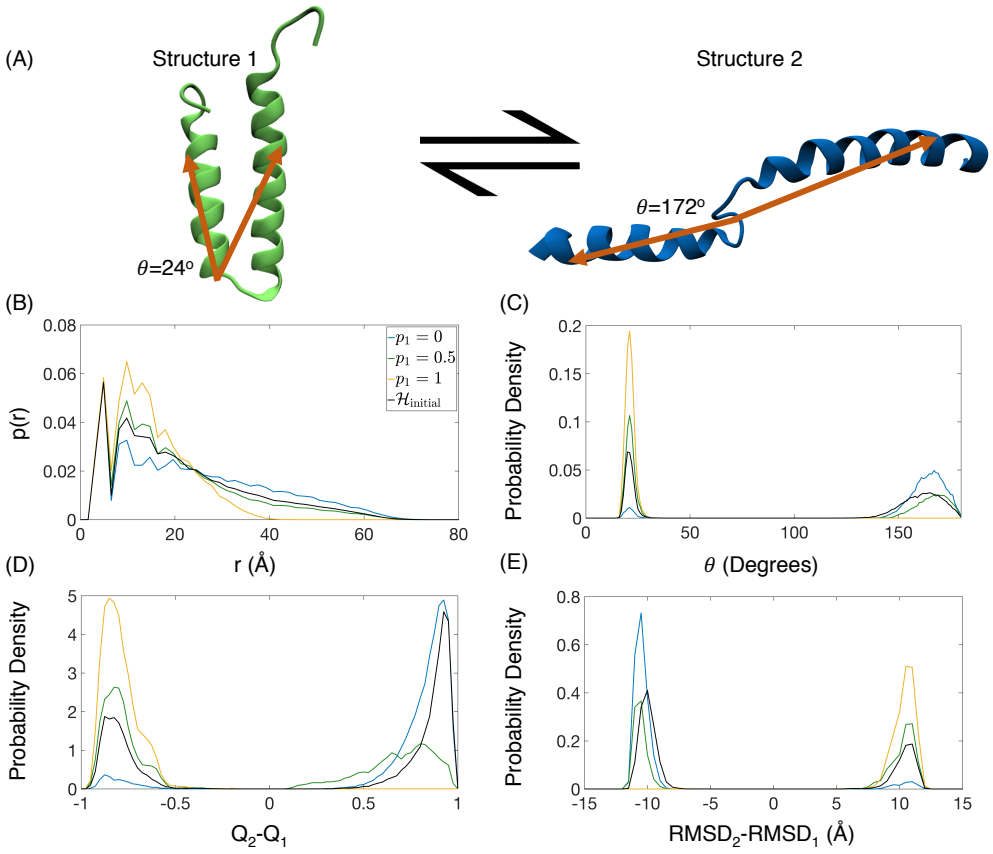


Figure 6: The maximum entropy approach succeeds in extracting the conformational population of a multi-domain protein from SAXS data. (A) Representative structures from the two energy basins adopted by the protein. (B) Comparison of the pair-wise distance distribution, $p(r)$, determined with the initial Hamiltonian $\mathcal{H}_{\text{initial}}$ (black), from structure 1 (yellow), from structure 2 (blue), and from an equal mix of both structures (green). (C, D, E) Probability distributions as a function of different structural variables calculated using $\mathcal{H}_{\text{initial}}$ (black) and using AWSEM-SAXS derived with $p_1 = 0$ (blue), $p_1 = 0.5$ (green) and $p_1 = 1$ (yellow).

Table 1: Difference between experimental and simulated scattering profiles.

| PDB ID | $\chi^2$ before bias | $\chi^2$ after bias | $\chi^2_{0.15}$ before bias | $\chi^2_{0.15}$ after bias |
|---|---|---|---|---|
| 3mzq | 3.154 | 1.124 | 5.077 | 0.894 |
| 1soy | 2.397 | 1.515 | 2.828 | 1.347 |

# TOC IMAGE

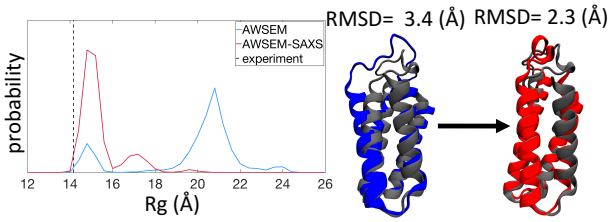


Figure 7: TOC Graphic