Article

Refining conformational ensembles of flexible proteins against small-angle X-ray scattering data

Francesco Pesce[1] and Kresten Lindorff-Larsen[1,*]

[1]Structural Biology and NMR Laboratory, The Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark
[*]Correspondence: lindorff@bio.ku.dk

ABSTRACT Intrinsically disordered proteins and flexible regions in multi-domain proteins display substantial conformational heterogeneity. Characterizing the conformational ensembles of these proteins in solution typically requires combining one or more biophysical techniques with computational modelling or simulations. Experimental data can either be used to assess the accuracy of a computational model or to refine the computational model to get a better agreement with the experimental data. In both cases, one generally needs a so-called forward model, i.e. an algorithm to calculate experimental observables from individual conformations or ensembles. In many cases, this involve one or more parameters that need to be set, and it is not always trivial to determine the optimal values or to understand the impact on the choice of parameters. For example, in the case of small-angle X-ray scattering (SAXS) experiments, many forward models include parameters that describe the contribution of the hydration layer and displaced solvent to the background-subtracted experimental data. Often, one also needs to fit a scale factor and a constant background for the SAXS data, but across the entire ensemble. Here, we present a protocol to dissect the effect of free-parameters on the calculated SAXS intensities, and to identify a reliable set of values. We have implemented this procedure in our Bayesian/Maximum Entropy framework for ensemble refinement, and demonstrate the results on three intrinsically disordered proteins and a three-domain protein connected by flexible linkers. Our results show that the resulting ensembles can depend substantially on the parameters used for solvent effects, and suggests that these should be chosen carefully. We also find a set of parameters that work robustly across all proteins.

SIGNIFICANCE The flexibility of a protein is often key to its biological function, yet understanding and characterizing its conformational heterogeneity is difficult. We here describe a robust protocol for combining small-angle X-ray scattering experiments with computational modelling to obtain a conformational ensemble. In particular, we focus on the contribution of protein hydration to the experiments and how this is included in modelling the data. Our resulting algorithm and software should make modelling intrinsically disordered proteins and multi-domain proteins more robust, thus aiding in understanding the relationship between protein dynamics and biological function.

INTRODUCTION

Small-angle X-ray scattering (SAXS) experiments are widely used in the field of integrative structural biology as a versatile tool to probe conformational ensembles of biomolecules in solution. When faced with highly dynamic and flexible systems, solving a crystallographic structure may either not be possible or only provide a static image that does not capture key aspects of the system. SAXS experiments instead give low to medium resolution ensemble-averaged view of the biomolecule. While for relatively rigid macromolecules it may be possible to derive a single averaged shape directly from a SAXS experiment (1–4), this is generally not possible for very flexible molecules. The possibility to calculate SAXS profiles from atomic coordinates, however, makes it possible to average across a distribution of conformations (a 'prior distribution'), compare the result with an experiment and to correct the prior distribution in case of poor agreement (5, 6).

SAXS measurements report on the total scattering of X-rays from all molecules in solution. Thus, the resulting scattering profile represents the entire solution (buffer and solute). For this reason, one also collects scattering data for the buffer alone, which is then subtracted from the data from the solution with the macromolecule to get the excess scattering. As the density of solvent around the protein (the hydration layer) may differ from that of the bulk solvent (7), the resulting data (colloquially termed SAXS data, though in practice it's a difference between two SAXS measurements) represents the signal coming from the protein together with its hydration envelope.

The calculation of SAXS intensities is done using a so-called forward model, *i.e.* an algorithm to predict an experimental observable from a structural model. Several forward models for SAXS exist and a key distinction is how they treat the scattering contribution from the protein hydration layer (5). In particular, to account for the hydration layer contribution, modelling approaches for SAXS data generally fall into two distinct categories.

In explicit solvent models (8–11), displaced water molecules and perturbed properties of the hydration envelope are explicitly taken into account in the calculation of scattering intensities. In particular, the scattering from the solute (protein) and hydration layer effects is estimated by explicit subtraction of the scattering calculated of the solvated protein and the solvent alone.

In contrast, in implicit solvent models (12–15) the contribution of the hydration layer and displaced volume to the scattering is modelled typically through one or more parameters that need to be set. The effects of the hydration layer can be modelled by a hydration shell of some width, Δ, and with excess density (and thus changed scattering compared to bulk solvent), $\delta\rho$. Typically, only the product $\Delta \cdot \delta\rho$ is important, and thus often Δ is fixed (e.g. at 3Å) and only $\delta\rho$ needs to be set (or fit). Similarly, implicit solvent models may include a parameter for the effective atomic radius (r_0), which both affects the overall displaced volume but also to some extent the contribution of the hydration layer.

While the explicit solvent strategy may provide a more realistic representation of the hydration layer and its contribution to the scattering data, it can be computationally expensive and requires a force field and water model that accurately models protein-water interactions. While it has been shown that the calculations on folded proteins are not particularly sensitive to the choice of water model (9, 16), there is more uncertainty about the best models for protein-water interactions for disordered proteins (16–19).

In many cases, one would want to use an implicit solvent strategy to calculate scattering data because of the smaller computational overhead. On the other hand, as described above these methods may require setting parameters that describe e.g. the protein's solvent envelope, and it is not always clear how best to determine these. Here we note that many forward models for experimental data, including implicit-solvent-based forward models for SAXS, have mostly been developed, parametrised and benchmarked using globular and folded protein structures. For folded proteins, free-parameters in forward models may be determined using SAXS data for proteins with known structures, or fitted for a given structural model. This approach, however, is difficult to apply to disordered proteins because for these there is not a well-defined reference structure from e.g. crystallography and there is uncertainty in computational methods for generating distributions of conformations (16, 19–21), both in terms of sampling efficiently all the possible conformations with the right probabilities or in terms of parametrisation. Similarly, it is not reasonable to fit these parameters independently to each structure because of the risk of substantial overfitting (9, 16), and because it is expected that the properties of the hydration layer will not depend fundamentally on the details of the structure. Finally, a key problem is that SAXS calculations are often used to construct or bias conformational ensembles, so that a procedure needs to be able to determine the free-parameters self-consistently together with the conformational ensemble.

Here, we focus our attention on these issues with implicit solvent calculations of scattering data for heterogeneous ensembles of conformations. We illustrate the effect of varying the parameters describing the hydration layer and displaced volume on ensemble refinement of intrinsically disordered proteins (IDPs) and a flexible multi-domain protein. We do so via a new, iterative and self-consistent strategy to select and optimise free-parameters in SAXS calculations while at the same time constructing a conformational ensemble to represent the data.

METHODS

Generating conformational ensembles of IDPs

We generated conformational ensembles for the polypeptide backbones of three IDPs using Flexible-Meccano (22). As the complexity of the ensembles may be influenced by the lenght of the protein, we generated larger ensembles for the longer proteins: Hst5 (24 residues, 10000 conformers), Sic1 (90 residues, 15000 conformers) and Tau (441 residues, 30000 conformers). We added side chains to the backbone structures generated by Flexible-Meccano using PULCHRA (23) with default settings.

Iterative Bayesian/Maximum-Entropy reweighting scheme

In integrative structural modelling, one approach is to use reweighting to refine probability distributions in order to improve the agreement between calculated averages and experimental values (24). Here we use the Bayesian/Maximum entropy (BME) reweighting procedure (25) that, by minimal modification of the prior distribution and taking into account the uncertainty in the experimental observable (σ_i), modifies the prior weights ω_j^0 to minimise the pseudo free-energy functional (24, 26):

$$\mathcal{L}(\omega_1 \cdots \omega_n) = \frac{m}{2}\chi^2_{\text{red}}(\omega_1 \cdots \omega_n) - \theta S_{\text{rel}}(\omega_1 \cdots \omega_n) \tag{1}$$

Here m is the number of experimental data points, $(\omega_1 \cdots \omega_n)$ are the weights associated to each conformer of the ensemble, the reduced χ^2 quantifies the agreement of the weighted average forward model predicted from each conformation x_j $(F(x_j))$ with the experimental data F_i^{EXP}:

$$\chi^2_{\text{red}}(\omega_1 \cdots \omega_n) = \frac{1}{m} \sum_i^m \frac{(\sum_j^n \omega_j F(x_j) - F_i^{EXP})^2}{\sigma_i^2} \tag{2}$$

$S_{\text{rel}} = -\sum_j^n \omega_j \ln \frac{\omega_j}{\omega_j^0}$ is the relative entropy that quantifies how much the reweighted distribution deviates from the prior. θ is a temperature-like free-parameter that effectively sets the balance between experiments and computation, and takes into account various sources of error such as inaccuracies in the force field or the forward model. In the limit $\theta \rightarrow$ inf, no confidence is assigned to the experimental data and no reweighting is performed. As θ is decreased, more weight is put on improving agreement with experiments, but at the cost of an increased deviation between the posterior (refined) distribution and the prior distribution (in this case generated by Flexible-Meccano or Martini). This can also be quantified as $\phi_{\text{eff}} = \exp(S_{\text{rel}})$, corresponding to the fraction of the original n frames that effectively contribute to the refined ensemble. For $\theta = 0$, the χ^2 is minimised without considering the prior distribution, in some cases leading to very low values of ϕ_{eff}. In methods such as BME, θ should be chosen in such a way to find the balance between minimising the χ^2 and retaining as much information as possible from the prior, such as for example when χ^2 reaches a plateau (24, 27).

Since the goal of the BME is to decrease the χ^2, it is important to ensure, when needed, that the experimental and calculated values are on the same scale. While SAXS intensities can be measured and calibrated on an absolute scale, this depends on careful calibration of the instrument and accurate measurements of the protein concentration. Thus, often calculated SAXS profiles are rescaled to match the experimental data. Moreover, experimental SAXS data may contain a small, non-zero background scattering, e.g. from imperfect background subtraction, which sometimes is dealt with by shifting calculated SAXS profiles to get a better fit.

To account for these issues, we here present the **iBME** (iterative Bayesian/Maximum-Entropy) approach, an iterative scheme that we have developed with the aim of coupling ensemble refinement and optimization of scale factor and constant background of the calculated SAXS profiles. The scheme is structured as follow:

1. Given a set of SAXS profiles calculated from each structure in a conformational ensemble, the corresponding ensemble-averaged SAXS profile is calculated using a set of initial (prior) weights (uniform weights in all our ensembles). We then perform a weighted least squares fit between the ensemble-averaged calculated SAXS profile and the experimental SAXS profile to get slope and intercept of the resulting linear fit. Weights for the weighted least squares fit are defined as $\frac{1}{(\sigma_i^2)}$.

2. The slope and intercept from (1) are used as scale factor and constant background to rescale and shift the calculated SAXS profiles.

3. BME is used for weight optimisation starting from the prior weights.

4. The optimised weights from (3) are used to calculate a new ensemble average of the SAXS profiles, which in turn is used for a new weighted least square fit to the experimental profile.

5. With the new slope and intercept, the calculated SAXS dataset used in the previous BME reweighting is again rescaled and shifted.

6. Repeat 3–5 until the drop of χ^2_{red} between consecutive iteration of the algorithm falls below a predefined threshold, or for a fixed number of iterations (we used 20 iterations in our analyses).

We initially tested the method using synthetic data to examine how well it can recover the scale factor and constant background (see SI text and Figs. S1–S4). We note also that iBME in part has overlap with features in BioEn (28), where only the scale factor is adjusted iteratively upon weight optimisation.

iBME is implemented in an updated version of BME (`https://github.com/KULL-Centre/BME`). Data and scripts used for the analyses presented in this manuscript are available at `https://github.com/KULL-Centre/papers/tree/master/2021/SAXS-pesce-et-al`.

Calculation of the radii of gyration

We use two different methods to estimate the average radius of gyration (R_g) of a conformational ensemble, one based on the protein coordinates and another based on the SAXS data.

From a conformational ensemble, the R_g for each conformer of n atoms can be calculated as $R_g = \sqrt{\frac{\sum_i^n m_i |r_i - r_{\mathrm{COM}}|}{\sum_i^n m_i}}$, with r_i being the position of the ith atom, m_i its mass and $r_{\mathrm{COM}} = \frac{\sum_i^n r_i m_i}{\sum_i^n m_i}$ in the centre of mass. We used MDTraj (29) for these calculations, and calculate the ensemble average, $\langle R_g \rangle$, as a linear or weighted average of the R_gs from each conformer.

From an experimental SAXS profile, we use the Guinier approximation to estimate the average R_g in solution (30). We first transform the SAXS profile as $\ln I(0)$ vs q^2, and then obtain $\langle R_g \rangle$ from the slope (a) of a linear fit in the small-angle region using $\langle R_g \rangle = \sqrt{-3a}$. The linear fit takes into account the uncertainty on the intensities (propagated as $|\frac{\sigma_i}{I_i}|$) and was performed using the scikit-learn python library (31).

RESULTS AND DISCUSSION

Conformational ensembles and SAXS data

Our aim here is to develop a strategy to model accurately conformational ensembles of flexible proteins with SAXS data, taking into account both uncertainty about a scale factor and constant background in the experimental SAXS data as well as effects of the hydration layer and displaced solvent. As object for our analyses we selected four proteins for which SAXS profiles had been determined experimentally and published. Also, protein flexibility may exist in multiple forms and to include different types, we choose three IDPs of different lengths and a multi-domain protein with flexible linkers: Histatin 5 (Hst5) (SAXS data collected at 323K from (32)), Sic1 (SAXS data from (33)), full-length (ht40-)Tau (SAXS data from (34)), and the three-domain protein TIA-1 without its flexible low-complexity domain (SAXS data from (35)).

We generated conformational ensembles of the three IDPs using Flexible-Meccano (22), while we used a previously generated molecular dynamics simulation of TIA-1 (36). The TIA-1 simulations were performed with the Martini force field (37) after increasing the interaction strength between protein and water by 6% (36). All structures were converted to all-atom representation before calculating SAXS data.

We used the implicit solvent SAXS-calculation approach Pepsi-SAXS (13) to calculate SAXS profiles from atomic coordinates. We choose this method for its versatility and computational efficiency, but our approach will also apply to other similar methods (12, 14). When no additional information, other than atomic coordinates and an experimental SAXS profile, is provided to Pepsi-SAXS, the software may tune four parameters in order to optimize the fit between the calculated and experimental SAXS profile: (i) the intensity of the forward scattering $I(0)$ (*i.e.* the scale of the profiles), (ii) a constant background cst, (iii) the effective atomic radius r_0 and (iv) the contrast of the hydration layer $\delta\rho$. In our calculations, we do not enable direct parameter fitting within Pepsi-SAXS and instead keep these parameters fixed to the same value for each conformer of an ensemble. As described in more detail below, we instead fit $I(0)$ and cst as global ensemble averages, and scan r_0 and $\delta\rho$ to determine self-consistent ensembles.

Determining self-consistent ensembles and hydration layer and displaced solvent parameters

By default, Pepsi-SAXS performs a grid search for the combination of r_0 and $\delta\rho$ that provides the best fit (lowest χ^2) between the SAXS profile calculated from of a specific protein structure and the experimental data. While this strategy may be appropriate to calculate a SAXS profile for globular proteins with little conformational heterogeneity, it can result in overfitting if applied to each structure in a highly heterogenous conformational ensembles. Default values of r_0 and $\delta\rho$ might be determined by fitting SAXS data to known crystal structures, and used without modification on other proteins. This, however, would amount to making the assumption that the hydration effects are constant and transferable from specific globular proteins to e.g. IDPs. We note here that it has been shown that the surface properties of the protein affect the hydration contribution (38, 39).

Instead, to determine the combination of parameters that best describes a conformational ensemble, to shed light on the influence of these two parameters, and to find a single set of parameters that provides a good description of the data, we here want to keep the *rationale* of a grid search, but adding an ensemble perspective. Similar to the standard grid scan, we calculate SAXS data for a range of values of r_0 and $\delta\rho$. Given that the input ensemble may not be accurate, we do not, however, simply score the grid based on the χ^2. Instead, for each combination of r_0 and $\delta\rho$ we perform an ensemble refinement using our newly described iBME approach (see Methods). To define the ranges for the grid, we compare the search ranges for r_0 and $\delta\rho$ implemented by three of the most widely used algorithms for SAXS calculations: CRYSOL (12), FoXS (14) and Pepsi-SAXS (Table S1), and use the widest ranges allowed by the three methods. Specifically, for r_0, we use 11 values in the range 1.4 – 1.8 Å, while for $\delta\rho$ we use 30 values in the range -27.0 – 70.0 e/nm^3. We also make the assumption that r_0 and $\delta\rho$ are the same for all conformers in the ensemble. While these might in principle be conformation dependent, we do so to decrease the risk of overfitting when varying these two parameters for each of the thousands of conformations.

Given the input ('prior') ensemble may not be fully representative of the protein in solution we do not just compare the

Table 1: Best fitting SAXS parameters, input and results of the iBME optimisation

	Hst5	Sic1	Tau	TIA-1
r_0 [Å]	1.722	1.558	1.640	1.722
$\delta\rho$ $[e/nm^3]$	10.02	10.02	0.00	-3.34
θ	80	80	50	100
χ^2_{red} (before iBME)	3.52	1.39	1.64	0.919
χ^2_{red} (after iBME)	1.04	1.02	1.14	0.540
ϕ_{eff}	0.911	0.941	0.707	0.884

experimental SAXS profiles with each of the average SAXS profiles calculated with Pepsi-SAXS with different values of r_0 and $\delta\rho$ (24, 40). Instead, we use the BME approach to reweight the prior ensembles against the experimental data (25), using as input (one at a time) the SAXS calculations with different values of r_0 and $\delta\rho$. In the reweighting it is key that the calculated SAXS profiles match the intensity of the forward scattering $I(0)$ and the constant background cst of the experimental signal. To accurately fit both $I(0)$ and cst upon reweighting, we developed the iBME scheme (see Methods for detailed description and SI for validation). This new method, uses iterations of rescaling and shifting the calculated SAXS profiles and reweighting the conformational ensemble to fit a global value of $I(0)$ and cst. The only requirement is that the same values for both $I(0)$ and cst are used to calculate the SAXS data for all conformers ($I(0)$ and cst are ensemble properties related to the experimental SAXS profile and independent on the single conformation). We set $I(0) = 1$ and $cst = 0$ for all conformers in the Pepsi-SAXS calculations, but as these parameters are adjusted by iBME, the choice of the starting values is not essential. In this way, we scan a range of r_0 and $\delta\rho$ and use iBME to fit $I(0)$, cst, and the conformational ensemble. To simplify interpretations and analysis, we kept the parameter θ constant for each protein (values specified in Table 1). This was done to keep the balance between the prior and experiment constant, so as to focus on changes that arise due to differences in the hydration and displaced solvent parameters. The resulting reweighted ensembles (at different values of r_0 and $\delta\rho$) are analysed further below.

A scoring function for the ensembles on the $r_0 \times \delta\rho$ grid

Once we calculated SAXS profiles and refined (i.e. reweighted) the ensembles for each pair of parameters on the $r_0 \times \delta\rho$ grid, we needed a scoring function to quantify the agreement with the experimental data. We already note here the complication arising from the fact that the ensembles have been refined against the experiments.

We first calculated the χ^2_{red} after the iBME optimization to indicate the quality of the ensembles. For all proteins this lead to a large region with low-χ^2_{red} (Fig. 1a-d-g-j), suggesting that most of the combinations of r_0 and $\delta\rho$ tested with $r_0 \leq 1.722$Å can can be fitted to the experimental data. We note also that although the χ^2_{red} is widely used for the purpose of comparing SAXS profiles, it has been noticed that it can be prone to overfitting is the noise is not estimated correctly (41, 42). For this reason, previous studies have focused e.g. on identifying the amount of information in a SAXS profile or in correcting the experimental noise (42–44). As we are here comparing different fits to the same data and with the same number of degrees of freedom, we did not use such corrections here. In turn this means that the calculated values of χ^2_{red} cannot easily be compared across the four systems.

When reweighting an ensemble against experiments, it is important to monitor the effective fraction of frames (ϕ_{eff}) that, as explained above, quantifies how much the posterior distribution deviates from the prior. When ϕ_{eff} is low this indicates that the ensemble has to be modified substantially to achieve the desired agreement with experiments. In order to ease comparison across the different ensembles in the grid we have chosen to use the same value of θ for all ensembles, where θ sets the balance between not deviating too much from the prior ensemble (maximizing ϕ_{eff}) and fitting the experimental data (minimizing χ^2_{red}). Thus at fixed values of θ the resulting value of ϕ_{eff} is another indicator of the quality of the ensembles (45), and we find a relative narrow region of the grids with high values of ϕ_{eff} (Fig. 1b,e,h,k). Thus comparing the maps of χ^2_{red} and ϕ_{eff} we find that while it is possible to achieve a relative good fit at a wider range of values of $\delta\rho$ and r_0, in many cases this comes at the cost of a substantial deviation from the prior (low ϕ_{eff}). To combine the balance of achieving a low χ^2_{red} and a high ϕ_{eff} we thus introduce a variable, $\gamma = \log(\frac{\chi^2_{\text{red}}}{\phi_{\text{eff}}})$, that combines these two effects in a single number (Fig. 1c,f,i,l). The results show that it is not possible to obtain a good fit (defined here as giving rise to a low γ) at all values of $\delta\rho$ and r_0, but that there are certain regions that appear to give rise to comparable fits. The parameter set that gives the lowest value of γ is reported in Table 1 together with the χ^2_{red} before and after reweighting and the ϕ_{eff}. We note that the final ensemble may not be optimal, and that further refinement could be obtained by scanning θ (24, 25, 27).

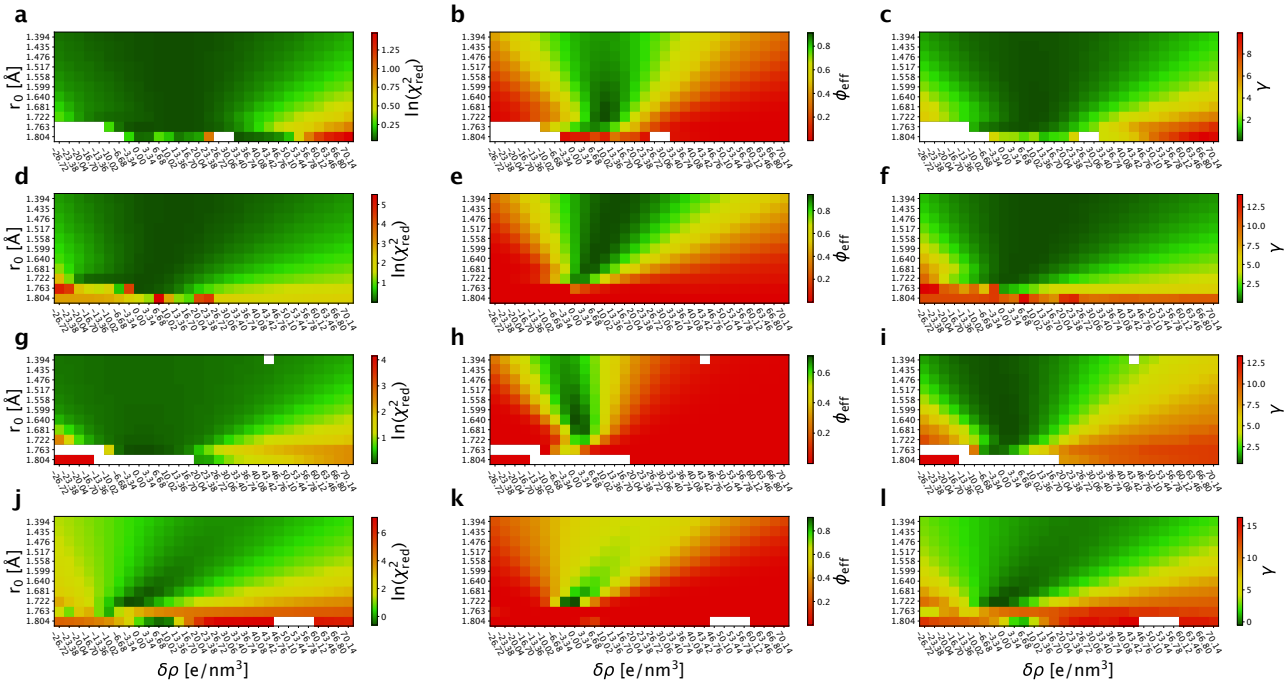

Figure 1: Reweighting ensembles using SAXS data calculated using different values for the parameters that effect the contribution from for the hydration layer and displaced solvent. The grids show the results from the iBME ensemble optimization with different combinations of $\delta\rho$ and r_0. The top row (a–c) show Hst5, the second row (d–f) show Sic1, the third row (g–i) show Tau, and the last row (j–l) show results for for TIA1. For each protein we show in the first column (a, d, g, j) $\ln\left(\chi^2_{\mathrm{red}}\right)$, the second column (b, e, h, k) ϕ_{eff}, and third column (c, f, i, l) $\gamma = \ln\left(\frac{\chi^2_{\mathrm{red}}}{\phi_{\mathrm{eff}}}\right)$. White spots correspond to ensembles where the iBME reweighting failed.

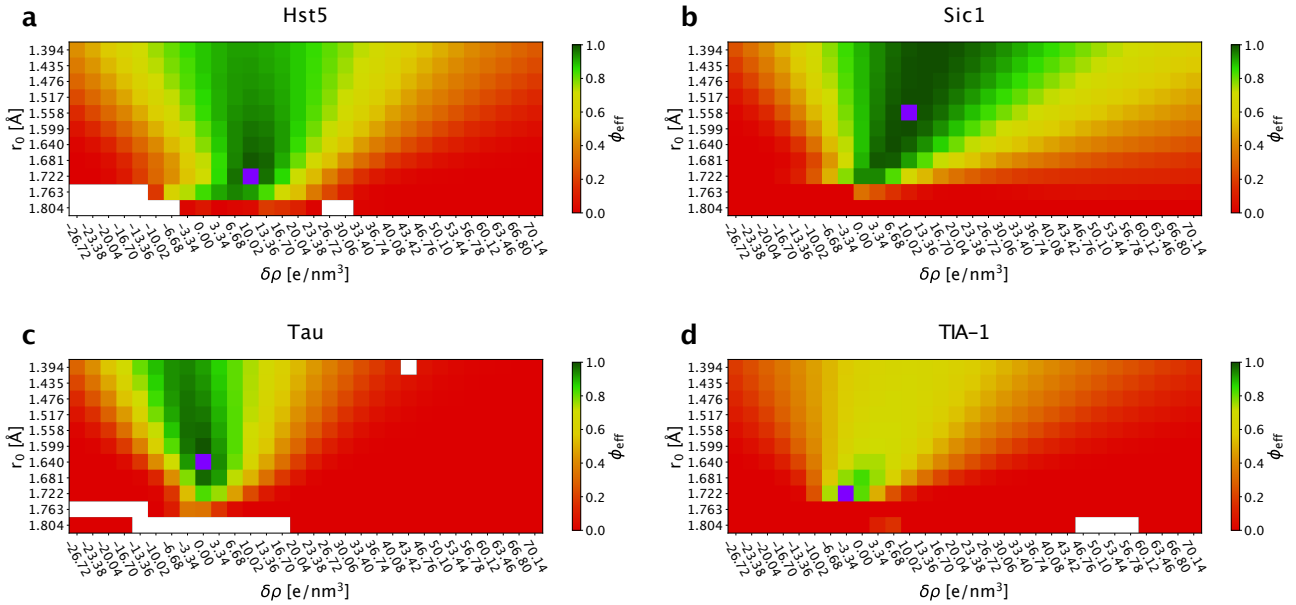

Figure 2: Comparing ensembles relative to the optimum. For each protein we calculated the effective fraction of frames (shown here as ϕ_{eff}) between the weights obtained using the parameters in Table 1 and the weights obtained at all other combinations of r_0 and $\delta\rho$. White spots correspond to ensembles where the BME reweighting failed.

We observe that while there are some differences between the four proteins, it also appears that there is a region that gives relative good fits for all proteins (Fig. 1). Because it may be computationally expensive to scan many sets of parameters, we also aimed to find a set of parameters that provides good scores for all four proteins. We therefore normalised and averaged the γ scores, and found the minimum to be at $\delta\rho = 3.34\ e/nm^3$ and $r_0 = 1.681$ Å.

Effect of hydration and atomic radius parameters on the conformational ensemble

The idea of the grid search is to find a combination of r_0 and $\delta\rho$ that gives rise to the best agreement with experimental data, taking also into account that we need to determine the parameters and ensemble weights at the same time. Here we explore the effect of choosing specific sets of these parameters on the conformational ensembles.

We first examined how much the individual ensembles differed from that determined using the r_0 and $\delta\rho$ parameters that give rise to the lowest value of γ (Table 1). We therefore calculated, as a measure of the difference between ensembles, ϕ_{eff} between the weights optimized using the different combinations of r_0 and $\delta\rho$ relative to the weights obtained using the 'optimal' values of r_0 and $\delta\rho$ (Fig. 2). As expected, values around the optimum give rise to comparable weights (ϕ_{eff} close to 1). For Sic1 and TIA-1 we also note a correlation between r_0 and $\delta\rho$, so that increasing the excess density ($\delta\rho$) and decreasing the radius (r_0) appears to give rise to more comparable ensembles. Nevertheless, the results also show that while several different combinations of r_0 and $\delta\rho$ can give rise to a good fit (Fig. 1), the resulting ensembles differ depending on the choice of parameters used to calculate the scattering data. In particular, we find that the ensembles are rather sensitive to the choice of $\delta\rho$, in particular for the three IDPs.

SAXS data are often used to estimate the radius of gyration (R_g), and so we demonstrate how the different ensembles have different distributions of R_g. Using Sic1 as an example we chose the optimal parameters as well as three other combinations of r_0 and $\delta\rho$ and calculated $p(R_g)$ after reweighting (Fig. 3). The results show that as long as r_0 and $\delta\rho$ are chosen within the range that gives a low value of γ the resulting distribution is relatively similar. On the other hand, if more extreme values for the r_0 and $\delta\rho$ parameters are chosen, the average R_g may differ substantially in the reweighted ensembles (Fig. S5).

Comparing ensembles to experimental estimates of R_g

In the analyses of the R_g described above, we implicitly referred to the values calculated from the protein coordinates as the mass-weighted root-mean-square distance from the centre of mass of the protein. This is a geometric quantity that is often used to study protein behaviour and biophysics. Because the ensembles were constructed by fitting to the experimental SAXS data,

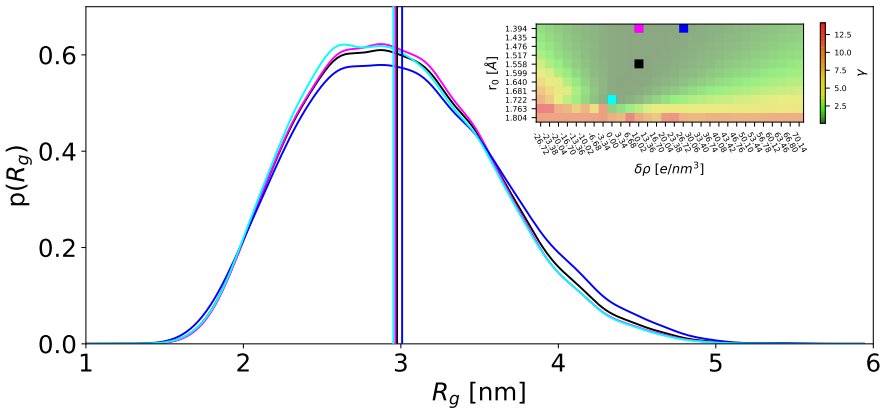

Figure 3: Effect of the $\delta\rho$ and r_0 parameters on reweighted probability distributions of R_g. We use Sic1 as an example and show $p(R_g)$ from both the optimal (lowest γ) parameters (black) as well as three other choices of r_0 and $\delta\rho$ in the low-γ region.

the resulting averages and distributions of R_g represent the experimental system, but exactly because the hydration effects were included the SAXS calculations, this means that these R_g values only represent the protein.

Another approach to estimate the average R_g ($\langle R_g \rangle$) from experiment is to fit the SAXS data directly without resorting to a conformational ensemble. The most common approach is to use the Guinier approximation (30), though other approaches exist (46–48). Because the SAXS data potentially contains a contribution from the hydration layer, the $\langle R_g \rangle$ estimated by a Guinier analysis (or similar methods) may in principle contain contributions from this (16). One complication of a Guinier analysis is to identify the linear part of the curve (the Guinier region), in particular because the first few low-q points of the scattering curve may often be noisy. As rule of thumb, the maximum scattering angle that can be used for the Guinier approximation satisfies the condition $q_{\max}\langle R_g \rangle < 1.3$ (1), but a threshold value of 0.9 has also been proposed for disordered systems (49).

Because both approaches to estimate $\langle R_g \rangle$ are commonly used, we here compare the two results. In addition to shedding light on differences, this analysis is also relevant because it is relatively common to compare $\langle R_g \rangle$-values calculated from simulations with values estimated from experiments, even though the two might differ due to effects of the hydration layer. We thus performed a Guinier analysis of the SAXS data for the four proteins, progressively extending the upper limit of the q-range and plotting R_g vs $q_{\max}R_g$ (49). As an indication of the quality of the linear fit we calculate χ^2_{red} (having a measure of the linearity helps in identifying the Guinier region of the SAXS profile), noting that error estimates and correlations between data points makes it difficult to interpret χ^2_{red}. The results show that in most cases the region around $q_{\max}R_g = 1.3$ is where the linear fit is best and the resulting R_g value is more stable (Fig. 4). The outlier is the smallest of the proteins, Hst5, where the stable and best fitting region is between 0.8 and 1.0 of qR_g, which agrees with (49). Returning to the question of how the R_g values estimated from the Guinier fit compares to the average R_g from the conformational ensembles with the lowest γ scores (horizontal black line in Fig. 4), we find that these are in a reasonable agreement (within 0.2 nm) with the values calculated from Guinier fits using $q_{\max}R_g = 1.3$.

CONCLUSIONS

SAXS experiments are widely used as source of structural information and are often integrated with computational methods to determine conformational ensembles. Generally, such approaches rely a forward model such as Pepsi-SAXS (13) to calculate SAXS data from one or more conformations and optimizes the structures or weights to improve agreement with experiments. While these approaches are very powerful, they are subject to uncertainty due to the choice of unknown parameters in the forward model. In principle these parameters can be 'integrated out' using Bayesian approaches (50, 51), though this can become computationally prohibitive for SAXS calculations. The aim of our work is thus to provide a robust protocol that estimates values for the relevant free parameters. In the context of SAXS these include the two parameters that determine the effects of the hydration layer and displaced volume ($\delta\rho$ and r_0), as well as a scale factor and constant background ($I(0)$ and $cst0$) that are often necessary to estimate.

We have developed and tested iBME as an extension to the weight estimation in BME to include a scale factor and constant background between experimental and calculated values. Importantly, the values are estimated as the globally best fitting parameters and are determined self-consistently with the weights of the ensembles. While we have here presented iBME

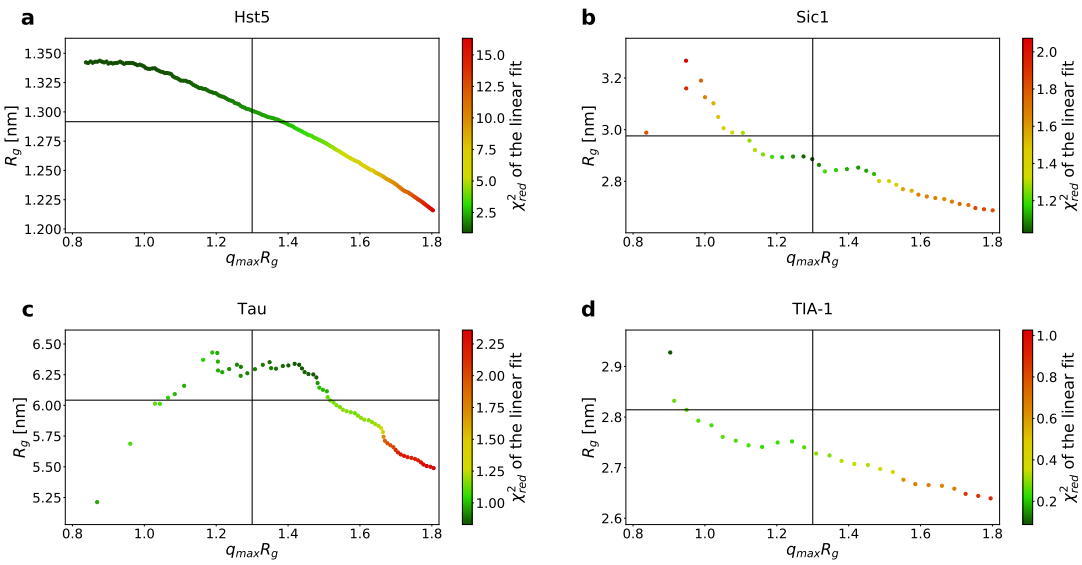

Figure 4: Estimating $\langle R_g \rangle$ from experiment using Guinier fitting and ensemble refinement. We used the Guinier approximation to estimate R_g by fitting from the lowest measured value of q to different values of $q_{\max}$, reporting the results as R_g vs. $q_{\max}R_g$. The colour indicates the χ^2_{red} of the linear fit. The vertical black lines represent the $q_{\max}R_g = 1.3$, and the horizontal black lines are the ensemble averaged R_g calculated from the conformational ensembles with the chosen optimal r_0 and $\delta\rho$ parameters (Table 1).

in the context of SAXS data, other types of data, such as NMR residual dipolar couplings, solvent paramagnetic relaxation enhancement or circular dichroism spectroscopy, may also involve estimating an overall scale. For small-angle neutron scattering data, the ability to include (fit) a constant background can be important because of contributions from incoherent scattering.

We also present the results from an extensive analysis of the effect of the r_0 and $\delta\rho$ parameters on calculated SAXS data and the resulting ensembles. We have determined self-consistent ensembles where the ensembles are reweighted using SAXS data calculated using different values for these parameters. Such an analysis is in particular important for large ensembles of flexible molecules, since fitting these parameters to each conformation could lead to substantial overfitting.

Combining these two aspects, the approach that we have described can be summarized as follows: (i) Sampling a conformational ensemble. (ii) Calculate SAXS profiles from the conformers of the ensemble, keeping scale parameters fixed ($I(0) = 1$, $cst = 0$), and perform a grid scan for $\delta\rho$ and r_0. (iii) For each value of $\delta\rho$ and r_0, optimise the weights, $I(0)$ and cst using iBME. (iv) Examine the results by calculating χ^2_{red}, ϕ_{eff} and γ, and select the ensemble with the lowest value of γ.

One complication of the algorithm is that it requires a large number of calculations of SAXS intensities. For this reason we also propose default values for $\delta\rho$ and r_0 that we find to provide relatively accurate results for all four proteins that we examined.

Finally, we also discuss considerations on the common practice of comparing the experimentally determined R_g (calculated with the Guinier approximation) with R_g calculated from the structural ensemble. While the results show good agreement, they also suggest that caution should be exerted when comparing average R_g values from simulations and experiments. In particular, we find that the both changing the r_0 and $\delta\rho$ parameters (Fig. S5) or the region used for Guinier fitting (Fig. 4), can change the R_g substantially, and so generally we recommend that it is better to compare the experimental data (in this case SAXS intensities) with values calculated from simulations, rather than comparing parameters estimated from experiments. Nevertheless, even such comparisons contain ambiguities as one needs to chose parameters in the SAXS calculations. We thus suggest that our work will be useful when benchmarking molecular simulations against SAXS data by providing additional insight into the effect of the hydration layer (16) and suggest default values that can be used as a starting point.

ACKNOWLEDGMENTS

We acknowledge Sandro Bottaro for fruitful discussions and for implementing the iBME algorithm in BME, Simone Orioli for useful discussions and input about SAXS and other aspects of this manuscript. We thank Andreas Haahr Larsen and Ramon

Crehuet for discussions and comments on the manuscript. We are grateful to Dimitri Svergun and Tanja Mittag for sharing SAXS data for, respectively, Tau and Sic1. This research was funded by the Lundbeck Foundation BRAINSTRUC initiative in structural biology (to KLL; R155-2015-2666).

REFERENCES

1. Kikhney, A. G., and D. I. Svergun, 2015. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters* 589:2570 – 2577. http://www.sciencedirect.com/science/article/pii/S0014579315007279, dynamics, flexibility, and intrinsic disorder in protein assemblies.

2. Grant, T. D., 2018. Ab initio electron density determination directly from solution scattering data. *Nature methods* 15:191–193.

3. Prior, C., O. R. Davies, D. Bruce, and E. Pohl, 2020. Obtaining tertiary protein structures by the ab initio interpretation of small angle X-ray scattering data. *Journal of chemical theory and computation* 16:1985–2001.

4. He, H., C. Liu, and H. Liu, 2020. Model reconstruction from small-angle x-ray scattering data using deep learning methods. *Iscience* 23:100906.

5. Hub, J. S., 2018. Interpreting solution X-ray scattering data using molecular simulations. *Current Opinion in Structural Biology* 49:18 – 26. http://www.sciencedirect.com/science/article/pii/S0959440X17300891, theory and simulation • Macromolecular assemblies.

6. Tria, G., H. D. T. Mertens, M. Kachala, and D. I. Svergun, 2015. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* 2:207–217. https://doi.org/10.1107/S205225251500202X.

7. Svergun, D. I., S. Richard, M. H. J. Koch, Z. Sayers, S. Kuprin, and G. Zaccai, 1998. Protein hydration in solution: Experimental observation by x-ray and neutron scattering. *Proceedings of the National Academy of Sciences* 95:2267–2272. https://www.pnas.org/content/95/5/2267.

8. Park, S., J. P. Bardhan, B. Roux, and L. Makowski, 2009. Simulated x-ray scattering of protein solutions using explicit-solvent models. *The Journal of Chemical Physics* 130:134114. https://doi.org/10.1063/1.3099611.

9. Chen, P.-c., and J. Hub, 2014. Validating Solution Ensembles from Molecular Dynamics Simulation by Wide-Angle X-ray Scattering Data. *Biophysical Journal* 107:435–447. https://linkinghub.elsevier.com/retrieve/pii/S0006349514006109.

10. Knight, C. J., and J. S. Hub, 2015. WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nucleic Acids Research* 43:W225–W230. https://doi.org/10.1093/nar/gkv309.

11. Köfinger, J., and G. Hummer, 2013. Atomic-resolution structural information from scattering experiments on macromolecules in solution. *Phys. Rev. E* 87:052712. https://link.aps.org/doi/10.1103/PhysRevE.87.052712.

12. Svergun, D., C. Barberato, and M. H. J. Koch, 1995. *CRYSOL* – a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *Journal of Applied Crystallography* 28:768–773. https://doi.org/10.1107/S0021889895007047.

13. Grudinin, S., M. Garkavenko, and A. Kazennov, 2017. *Pepsi-SAXS*: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallographica Section D* 73:449–464. https://doi.org/10.1107/S2059798317005745.

14. Schneidman-Duhovny, D., M. Hammel, and A. Sali, 2010. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Research* 38:W540–W544. https://doi.org/10.1093/nar/gkq461.

15. Schneidman-Duhovny, D., M. Hammel, J. Tainer, and A. Sali, 2013. Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophysical Journal* 105:962–974. https://www.sciencedirect.com/science/article/pii/S0006349513008059.

16. Henriques, J., L. Arleth, K. Lindorff-Larsen, and M. Skepö, 2018. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *Journal of Molecular Biology* 430:2521–2539. https://www.sciencedirect.com/science/article/pii/S0022283618301232, intrinsically Disordered Proteins.

17. Best, R. B., W. Zheng, and J. Mittal, 2014. Balanced Protein–Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation* 10:5113–5124. https://doi.org/10.1021/ct500569b, pMID: 25400522.

18. Piana, S., A. G. Donchev, P. Robustelli, and D. E. Shaw, 2015. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *The Journal of Physical Chemistry B* 119:5113–5123. https://doi.org/10.1021/jp508971m, pMID: 25764013.

19. Henriques, J., C. Cragnell, and M. Skepö, 2015. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment. *Journal of Chemical Theory and Computation* 11:3420–3431. https://doi.org/10.1021/ct501178z, pMID: 26575776.

20. Rauscher, S., V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmüller, 2015. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *Journal of Chemical Theory and Computation* 11:5513–5524. https://doi.org/10.1021/acs.jctc.5b00736, pMID: 26574339.

21. Palazzesi, F., M. K. Prakash, M. Bonomi, and A. Barducci, 2015. Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States. *Journal of Chemical Theory and Computation* 11:2–7. https://doi.org/10.1021/ct500718s, pMID: 26574197.

22. Ozenne, V., F. Bauer, L. Salmon, J.-r. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay, and M. Blackledge, 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 28:1463–1470. https://doi.org/10.1093/bioinformatics/bts172.

23. Rotkiewicz, P., and J. Skolnick, 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of Computational Chemistry* 29:1460–1465. https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20906.

24. Orioli, S., A. H. Larsen, S. Bottaro, and K. Lindorff-Larsen, 2020. Chapter Three - How to learn from inconsistencies: Integrating molecular simulations with experimental data. *In* B. Strodel, and B. Barz, editors, Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly, Academic Press, volume 170 of *Progress in Molecular Biology and Translational Science*, 123–176. https://www.sciencedirect.com/science/article/pii/S1877117319302121.

25. Bottaro, S., T. Bengtsen, and K. Lindorff-Larsen, 2020. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach, Springer US, New York, NY, 219–240. https://doi.org/10.1007/978-1-0716-0270-6_15.

26. Hummer, G., and J. Köfinger, 2015. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics* 143:243150. https://doi.org/10.1063/1.4937786.

27. Różycki, B., Y. C. Kim, and G. Hummer, 2011. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. *Structure* 19:109 – 116. http://www.sciencedirect.com/science/article/pii/S0969212610003953.

28. Hummer, G., and J. Köfinger, 2015. Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics* 143:243150. https://doi.org/10.1063/1.4937786.

29. McGibbon, R. T., K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, 2015. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* 109:1528 – 1532.

30. Guinier, A., 1939. La diffraction des rayons X aux très petits angles : application à l'étude de phénomènes ultramicroscopiques. *Annales de Physique* 11:161–237.

31. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

32. Jephthah, S., L. Staby, B. B. Kragelund, and M. Skepö, 2019. Temperature Dependence of Intrinsically Disordered Proteins in Simulations: What are We Missing? *Journal of Chemical Theory and Computation* 15:2672–2683. https://doi.org/10.1021/acs.jctc.8b01281, pMID: 30865820.

33. Gomes, G.-N. W., M. Krzeminski, E. W. Martin, T. Mittag, T. Head-Gordon, J. D. Forman-Kay, and C. C. Gradinaru, 2020. Integrating multiple experimental data to determine conformational ensembles of an intrinsically disordered protein. *bioRxiv* https://www.biorxiv.org/content/early/2020/02/27/2020.02.05.935890.

34. Mylonas, E., A. Hascher, P. Bernadó, M. Blackledge, E. Mandelkow, and D. I. Svergun, 2008. Domain Conformation of Tau Protein Studied by Solution Small-Angle X-ray Scattering. *Biochemistry* 47:10345–10353. https://doi.org/10.1021/bi800900d, pMID: 18771286.

35. Sonntag, M., P. K. A. Jagtap, B. Simon, M.-S. Appavou, A. Geerlof, R. Stehle, F. Gabel, J. Hennig, and M. Sattler, 2017. Segmental, Domain-Selective Perdeuteration and Small-Angle Neutron Scattering for Structural Analysis of Multi-Domain Proteins. *Angewandte Chemie International Edition* 56:9322–9325. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201702904.

36. Larsen, A. H., Y. Wang, S. Bottaro, S. Grudinin, L. Arleth, and K. Lindorff-Larsen, 2020. Combining molecular dynamics simulations with small-angle X-ray and neutron scattering data to study multi-domain proteins in solution. *PLOS Computational Biology* 16:1–29. https://doi.org/10.1371/journal.pcbi.1007870.

37. Monticelli, L., S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, 2008. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *Journal of Chemical Theory and Computation* 4:819–834. https://doi.org/10.1021/ct700324x, pMID: 26621095.

38. Virtanen, J. J., L. Makowski, T. R. Sosnick, and K. F. Freed, 2010. Modeling the Hydration Layer around Proteins: HyPred. *Biophysical Journal* 99:1611–1619. https://www.sciencedirect.com/science/article/pii/S0006349510007393.

39. Virtanen, J., L. Makowski, T. Sosnick, and K. Freed, 2011. Modeling the Hydration Layer around Proteins: Applications to Small- and Wide-Angle X-Ray Scattering. *Biophysical Journal* 101:2061–2069. https://www.sciencedirect.com/science/article/pii/S0006349511010800.

40. van Gunsteren, W. F., X. Daura, N. Hansen, A. E. Mark, C. Oostenbrink, S. Riniker, and L. J. Smith, 2018. Validation of Molecular Simulation: An Overview of Issues. *Angewandte Chemie International Edition* 57:884–902. https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.201702945.

41. Rambo, R. P., and J. A. Tainer, 2013. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* 496:477–481. https://www.nature.com/articles/nature12070.

42. Larsen, A. H., and M. C. Pedersen, 2020. Experimental noise in small-angle scattering can be assessed and corrected using the Bayesian Indirect Fourier Transformation.

43. Hansen, S., 2000. Bayesian estimation of hyperparameters for indirect Fourier transformation in small-angle scattering. *Journal of applied crystallography* 33:1415–1421.

44. Konarev, P. V., and D. I. Svergun, 2015. *A posteriori* determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems. *IUCrJ* 2:352–360. https://doi.org/10.1107/S2052252515005163.

45. Qian, H., 2001. Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Physical Review E* 63:042103.

46. Vestergaard, B., and S. Hansen, 2006. Application of Bayesian analysis to indirect Fourier transformation in small-angle scattering. *Journal of applied crystallography* 39:797–804.

47. Riback, J. A., M. A. Bowman, A. M. Zmyslowski, C. R. Knoverek, J. M. Jumper, J. R. Hinshaw, E. B. Kaye, K. F. Freed, P. L. Clark, and T. R. Sosnick, 2017. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* 358:238–241.

48. Zheng, W., and R. B. Best, 2018. An extended Guinier analysis for intrinsically disordered proteins. *Journal of molecular biology* 430:2540–2553.

49. Borgia, A., W. Zheng, K. Buholzer, M. B. Borgia, A. Schüler, H. Hofmann, A. Soranno, D. Nettels, K. Gast, A. Grishaev, R. B. Best, and B. Schuler, 2016. Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *Journal of the American Chemical Society* 138:11714–11726. https://doi.org/10.1021/jacs.6b05917, pMID: 27583570.

50. Rieping, W., M. Habeck, and M. Nilges, 2005. Inferential structure determination. *Science* 309:303–306.

51. Spill, Y., Y. Karami, P. Maisonneuve, N. Wolff, and M. Nilges, 2021. Automatic Bayesian Weighting for SAXS data. *Frontiers in Molecular Biosciences* 8:495.