Biophysical Journal
Article

Biophysical Society

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

ABSTRACT Intrinsically disordered proteins and flexible regions in multidomain proteins display substantial conformational heterogeneity. Characterizing the conformational ensembles of these proteins in solution typically requires combining one or more biophysical techniques with computational modeling 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 involves 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 the 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 four intrinsically disordered proteins and a protein with three domains connected by flexible linkers. Our results show that the resulting ensembles can depend on the parameters used for solvent effects and suggest 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. Here, we describe a robust protocol for combining small-angle x-ray scattering experiments with computational modeling to obtain a conformational ensemble. In particular, we focus on the contribution of protein hydration to the experiments and how this is included in modeling the data. Our resulting algorithm and software should make modeling intrinsically disordered proteins and multidomain 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 a low-to-medium resolution ensemble-averaged view of the biomolecule. Whereas 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. Because the density of solvent around the protein

Submitted May 31, 2021, and accepted for publication October 4, 2021.
*Correspondence: lindorff@bio.ku.dk
Editor: Jill Trewhella.
https://doi.org/10.1016/j.bpj.2021.10.003

(the hydration layer) may differ from that of the bulk solvent (7), the resulting data (colloquially termed SAXS data, although in practice it is a difference between two SAXS measurements) represents the signal coming from the protein together with its hydration envelope and the solvent displaced by the protein.

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, modeling approaches for SAXS data generally fall into two distinct categories.

In explicit solvent models (8–12), 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 from the solvated protein and the solvent alone.

In contrast, in implicit solvent models (13–16), the contribution of the hydration layer and displaced volume to the scattering is modeled typically through one or more parameters that need to be set. The effects of the hydration layer can be modeled 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 $\Delta \times \delta\rho$ is important, thus Δ is often fixed (e.g., at 3 Å), 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.

Whereas 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. Although it has been shown that the calculations on folded proteins are not particularly sensitive to the choice of water model (9,17), there is more uncertainty about the best models for protein-water interactions for disordered proteins (17–20).

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, and as described above, these methods may require setting parameters that describe, for example, the protein's solvent envelope, and it is not always clear how best to determine these. Here, we note that many forward models, including implicit solvent-based forward models for SAXS, have mostly been developed, parametrized, 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, for example, crystallography and there is uncertainty in computational methods for generating distributions of conformations (17,20–22), both in terms of sampling efficiently all the possible conformations with the right probabilities or in terms of parametrization. Similarly, it is not reasonable to fit these parameters independently to each structure because of the risk of substantial overfitting (9,17) 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 multidomain protein. We do so via an iterative and self-consistent strategy to select and optimize free parameters in SAXS calculations while at the same time constructing a conformational ensemble to represent the data.

Our approach is based on a reweighting approach that is rooted in Bayesian inference (23–31) and the maximum entropy principle (32–37). Although these methods show similarities to other approaches based, for example, on genetic algorithms (6,38,39) or Monte Carlo processes (40,41), they differ in how they balance prior information (often encoded in a force field) with the experimental data. This balance can be particularly important for disordered proteins in which the solutions are typically severely underdetermined and in which large ensembles are required to provide a realistic structural description of the conformations present in solution (38). Finally, we note that in some cases it is possible to absorb some effects of protein dynamics into the forward model rather than to represent it explicitly in the form of a conformational ensemble, and such modifications exist both for various types of NMR data (42–45) and x-ray scattering or diffraction data (46–49). Although this may be useful when studying small fluctuations around a well-defined "average" conformation or when the dynamics is of rigid bodies in a crystal, here we examine systems in which we explicitly represent a broad conformational ensemble.

MATERIALS AND METHODS

Generating conformational ensembles of IDPs

We generated conformational ensembles for the polypeptide backbones of four IDPs using flexible-meccano (50). Flexible-meccano implicitly represents a potential energy function derived from the populations of backbone dihedrals in loop regions in folded protein structures. The backbone chains are built by random sampling these potentials. Other methods exist to efficiently generate conformational ensembles of IDPs, also with the possibility of taking into account transient secondary structure elements (if part of the sequence is known to assume these) (51). We chose flexible-meccano for most of the analyses presented here because it has been shown to generate conformational ensembles of IDPs that are in good agreement with both NMR observables and SAXS data (52–55) without the need to provide any prior knowledge about the system. Because the complexity of the ensembles may be influenced by the length of the protein, we generated larger ensembles for the longer proteins, including Hst5 (24 residues, 10,000 conformers), Sic1 (90 residues, 15,000 conformers), α-Synuclein (140 residues, 20,000 conformers), and Tau (441 residues, 30,000 conformers). We added side chains to the backbone structures generated by flexible-meccano using PULCHRA (56) with default settings.

Iterative Bayesian/maximum entropy reweighting scheme

In integrative structural modeling, one approach is to use reweighting to refine probability distributions to improve the agreement between calculated averages and experimental values (37). Here, we use the Bayesian/maximum entropy (BME) reweighting procedure (36) 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 minimize the pseudo-free energy functional (24,26,37):

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

Here, m is the number of experimental data points, $(\omega_1 \cdots \omega_n)$ are the weights associated with 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}as follows:

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

and $S_{\mathrm{rel}} = -\sum_j^n \omega_j \ln \frac{\omega_j}{\omega_j^0}$ is the relative entropy that quantifies how much the reweighted distribution deviates from the prior.

Thus, when minimizing L, we aim to lower χ^2_{red} (to improve agreement with experiment), while not decreasing the relative entropy term too much, in such a way as to obtain the minimal modification of the prior distribution that results in a better agreement with the experimental data. The parameter 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 (24,37). In the limit $\theta \rightarrow \infty$, no confidence is assigned to the experimental data, and no reweighting is performed. Because θ is decreased, more weight is put on improving agreement with experiments (χ^2_{red}) 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 simulations with the Martini or all-atom force fields). This can also be quantified as $\phi_{\mathrm{eff}} = \exp(S_{\mathrm{rel}})$, corresponding to the fraction of the original n frames that effectively contributes to the refined ensemble. For $\theta = 0$, the χ^2 is minimized without considering the prior distribution, in some cases leading to very low values of ϕ_{eff}, and so very few conformations contribute to the final average. In methods such as BME, θ should be chosen in such a way as to find the balance between minimizing the χ^2 and retaining as much information as possible from the prior, such as, for example, when χ^2 reaches a plateau (24,37).

We highlight that, as a consequence of the points above, the prior distribution is an important part of the procedure because the goal of BME and related approaches is to perform the minimal modification of the prior to get a reasonable agreement with the experimental data. The definition of the prior weights is strictly dependent on the method used to sample conformations. In case of standard molecular dynamics simulations, in which the conformations are sampled from a Boltzmann distribution, or flexible-meccano, in which the conformations are sampled from specific backbone dihedral angle potential wells, the weights are uniform because the probability of a certain conformation is related to its occurrence in the ensemble.

Because 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. Whereas 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, calculated SAXS profiles are often rescaled to match the experimental data. Moreover, experimental SAXS data may contain a small, nonzero 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 present here the iterative Bayesian/maximum entropy (iBME) 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 follows:

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 optimizing the weights starting from the prior weights.
4) The optimized 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 data set 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 Supporting materials and methods and Figs. S1–S4). We note also that iBME, in part, has overlap with features in BioEn (26), in which only the scale factor is adjusted iteratively upon optimizing the weights.

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/main/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 i^{th} atom, m_i its mass, and $r_{\mathrm{COM}} = \frac{\sum_i^n r_i m_i}{\sum_i^n m_i}$ the center of mass. We used MDTraj (57) for these calculations and calculate the ensemble average $\langle R_g \rangle$ as a linear or weighted average of the R_g-values from each conformer.

As an alternative to using the atomic masses to weigh the distances in the calculation of R_g, we also use the atom contrasts, defined for the i^{th} atom as $\delta\rho_i^2 = (\rho_i - \rho_w)^2$, where ρ_w is the density of bulk water (334 e/nm^3) and ρ_i is the density of the i^{th} atom calculated as the ratio between its number of electrons and its volume (58). We do not, however, observe substantial differences between the mass-weighted and contrast-weighted values of R_g (Fig. S5).

From an experimental SAXS profile, we use the Guinier approximation to estimate the average R_g in solution (59). We first transform the SAXS profile as 1n $I(q)$ vs. q^2, 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 of the intensities (propagated as $\left|\frac{\sigma_i}{I_i}\right|$) and was performed using the scikit-learn python library (60).

RESULTS AND DISCUSSION

Conformational ensembles and SAXS data

Our aim here is to develop a strategy to model 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 the object for our analyses, we selected five proteins for which SAXS profiles had been determined experimentally and published. Also, because protein flexibility may exist in multiple forms and to include different types, we first choose three IDPs of different lengths and a multidomain protein with flexible linkers, including Histatin 5 (Hst5) (SAXS data collected at 323K from (61)), Sic1 (SAXS data from (62)), full-length (ht40-)Tau (SAXS data from (63)), and the three-domain protein TIA1 without its flexible low-complexity domain (SAXS data from (64)). Furthermore, we also analyze below an additional IDP (α-Synuclein, with SAXS data from (65)) to examine the robustness of the analyses done on the four proteins listed above.

We generated conformational ensembles of the four IDPs using flexible-meccano (50). Additionally we also analyzed two previously performed molecular dynamics simulations of α-Synuclein (66) produced using either the Amber a99SB-*disp* or the Amber ff03ws force field. We also used a previously generated molecular dynamics simulation of TIA1 (67). The TIA1 simulations were performed with the Martini force field (68) after increasing the interaction strength between protein and water by 6% (67). All structures were converted to all-atom representation before calculating SAXS data.

We used the implicit solvent SAXS calculation approach Pepsi-SAXS (Polynomial Expansions of Protein Structures and Interactions SAXS) (14) 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 (13,15), and below we also discuss and show results using FoXS. When no additional information, other than atomic coordinates and an experimental SAXS profile, is provided to Pepsi-SAXS, the software may tune four parameters to optimize the fit between the calculated and experimental SAXS profile: 1) the intensity of the forward scattering $I(0)$ (i.e., the scale of the profiles), 2) a constant background cst, 3) the effective atomic radius r_0, and 4) 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. Although 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 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, for example, IDPs. We note here that it has been shown that the surface properties of the protein affect the hydration contribution (69,70).

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, here, we want to keep the rationale of a grid search but add an ensemble perspective. Similar to the standard grid scan, we calculate SAXS data for a range of values of r_0 and $\delta\rho$. 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 (13), FoXS (15), 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, whereas for $\delta\rho$, we use 30 values in the range -27.0 to 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. Whereas these might in principle be conformation dependent (70,71), we do so to decrease the risk of overfitting when varying these two parameters for each of the thousands of conformations. Also, because the goal here is to describe the conformational distribution of the protein in solution, we do not expect a substantial difference as long as there is not a strong conformational dependency on the properties of the solvation layer.

Given that the input ("prior") ensemble may not be fully representative of the protein in solution, we do not just compare the experimental SAXS profiles with each of the average SAXS profiles calculated with Pepsi-SAXS with different values of r_0 and $\delta\rho$ (37,72). Instead, we use the BME approach to reweight the prior ensembles against the experimental data (36), 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 Materials and methods for detailed description and Supporting materials and methods for validation). The iBME method uses iterations of rescaling and shifting the calculated SAXS profiles and reweighting of 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 of the single conformation). We set $I(0)$ and $cst = 0$ for all conformers in the Pepsi-SAXS calculations, but because 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 because of differences in the hydration and displaced solvent parameters. The resulting reweighted ensembles (at different values of r_0 and $\delta\rho$) are analyzed further below.

We also validated the grid-scanning approach using synthetic SAXS data generated using a specific choice of r_0 and $\delta\rho$ to generate the data (see Supporting materials and methods). The results show that, both with a correct prior (Fig. S6) and a prior that is different from that used to generate the synthetic data (Fig. S7), the method is able to recover values of r_0 and $\delta\rho$ very close to those used to generate the data.

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.

TABLE 1 Best fitting SAXS parameters, input, and results of the iBME optimization

	Hst5	Sic1	Tau	TIA1
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

We first calculated the χ^2_{red} after the iBME optimization to indicate the quality of the ensembles. For Hst5, Sic1, Tau, and TIA1, this led to a large region with lowχ^2_{red} (Fig. 1, *a–j*), suggesting that most of the combinations of r_0 and $\delta\rho$ tested with $r_0 \leq 1.722$ Å 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 if the noise is not estimated correctly (73,74). For this reason, previous studies have focused, for example, on identifying the amount of information in a SAXS profile or in correcting the experimental noise (74–76). Here, because we are comparing different fits to the same data and with the same number of degrees of freedom, we did not use such corrections. In turn, this means that the calculated values of χ^2_{red} cannot easily be compared across the four systems that we analyzed.

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. To ease comparison across the different ensembles in the grid, we have chosen to use the same value of θ for all combinations of r_0 and $\delta\rho$, 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 (77), and we find a relatively narrow region of the grids with high values of ϕ_{eff} (Fig. 1, *b*, *e*, *h*, and *k*). Thus, comparing the maps of χ^2_{red} and ϕ_{eff}, we find that, whereas it is possible to achieve a relatively 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, $\gamma = \ln\left(\frac{\chi^2_{red}}{\phi_{eff}}\right)$, that combines these two effects in a single number (Fig. 1, *c*, *f*, *i*, and *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 sets that give the lowest values of γ for Hst5, Sic1, Tau, and TIA1 are 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 (24,36,37).

We observe that, whereas there are some differences between the four proteins analyzed above, it also appears that there is a region that gives relatively good fits for all

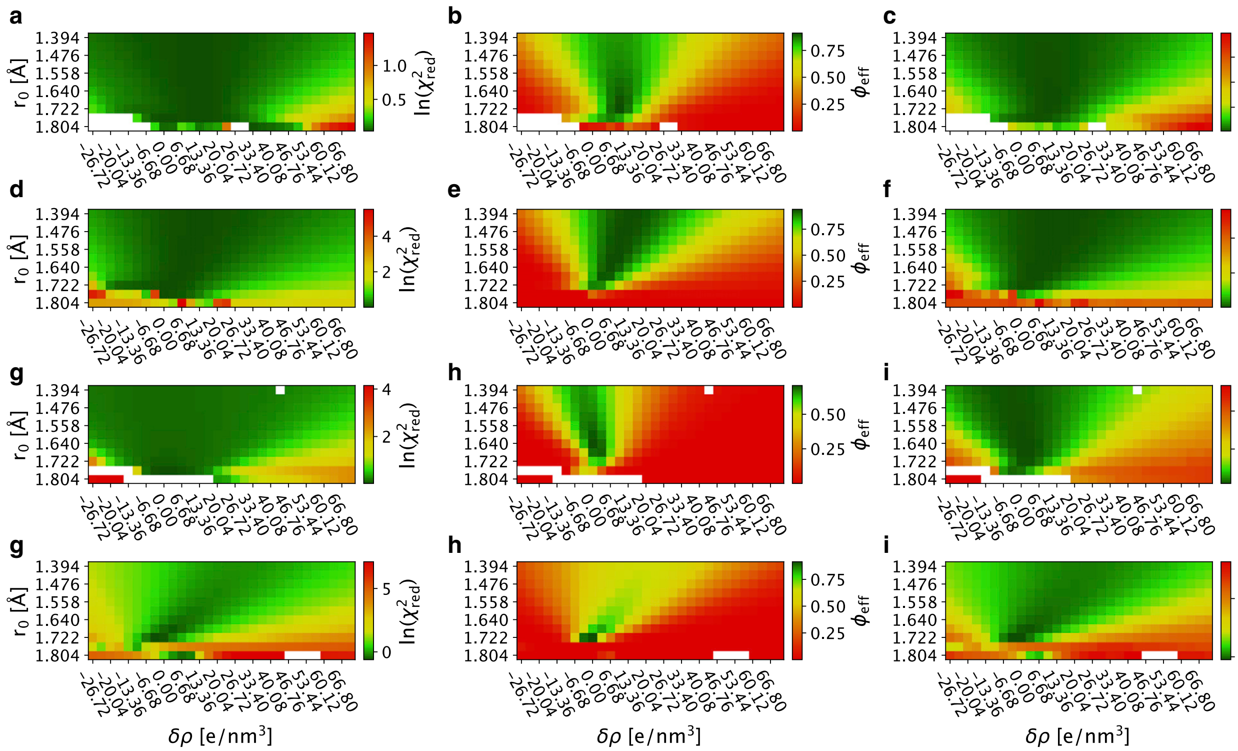


FIGURE 1 Reweighting ensembles using SAXS data calculated using different values for the parameters that effect the contribution from for the hydration layer and displaced solvent. The grids show the results from the iBME ensemble optimization with different combinations of $\delta\rho$ and r_0. The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

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 these four proteins. We therefore normalized 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.

We note that, in the literature, higher values are generally reported as default for $\delta\rho$ (generally 10% (7) or 6% (17,71) of the bulk density). In the context of SAXS calculations, however, we also note that the main quantity that determines the contribution of the hydration layer is the product between $\delta\rho$ and its width Δ (13,14). Whereas Δ is 3 Å in CRYSOL, it is chosen in a slightly different fashion in Pepsi-SAXS, and it is 5 Å in most of the cases that we examined. To demonstrate that different values for the contrast of the hydration layer alone can lead to the same result when the width is treated in different ways, we also repeated the grid scans for Hst5, Sic1, Tau, and TIA1 employing FoXS. Although the minima for γ are different from those obtained using Pepsi-SAXS (Fig. S8), the reweighted distributions of R_g from these minima are essentially identical (Fig. S9), reinforcing the observation that $\delta\rho$ alone is meaningful only in the context of a specific SAXS calculator. In addition, by again normalizing and averaging the γ score, we obtain a global minimum for the parameters in FoXS at r_0 =1.68 Å (as in Pepsi-SAXS) and $\delta\rho = -7.07$ e/nm^3. We note that this value for $\delta\rho$ appears substantially different from those used for folded proteins and suggest that further studies are needed to examine better the physical origins of these effects.

Conversely, the value $r_0 = 1.681$ Å is slightly higher than the average values used by CRYSOL and FoXS (1.62 Å) and Pepsi-SAXS (1.64 Å). Although the origin of this observation is unclear, we note that there are differences in protein volume, depending on whether a protein is folded or unfolded (78,79). Thus, the excluded volume, as described by the Fraser model (58), might need different parameters for compact and expanded proteins, and we suggest that this could be studied further using molecular simulations (9).

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, also taking into account that we need to determine

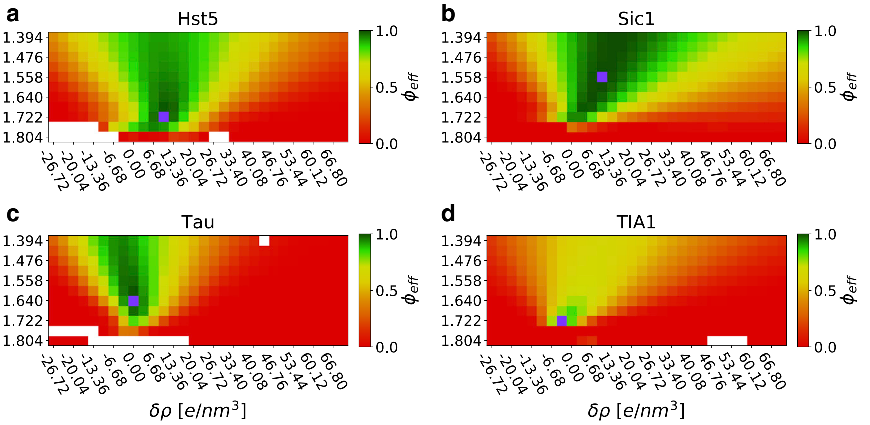


FIGURE 2 Comparing ensembles relative to the optimum. For each protein (*a*: Hst5, *b*: Sic1, *c*: Tau and *d*: TIA1) we calculated the effective fraction of frames (shown here as $\phi_{\rm eff}$) between the weights obtained using the parameters in Table 1 and the weights obtained at all other combinations of r_0 and $\delta\rho$. White spots correspond to ensembles in which the iBME reweighting failed. Purple spots correspond to the minima for γ. To see this figure in color, go online.

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.

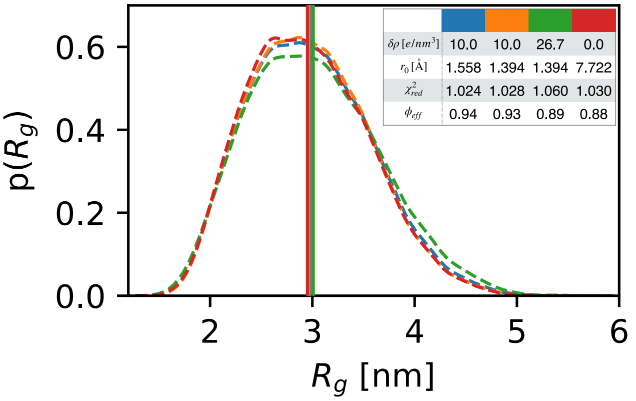
We first examined how much the individual ensembles differed from those determined using the r_0 and $\delta\rho$ parameters that give rise to the lowest value of γ (Table 1). We therefore calculated, as a measure of the difference between ensembles, $\phi_{\rm eff}$ between the weights optimized using 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 ($\phi_{\rm eff}$ close to 1). For Sic1 and TIA1, we also note a correlation between r_0 and $\delta\rho$, such that increasing the excess density ($\delta\rho$) and decreasing the radius (r_0) appear to give rise to more comparable ensembles. Nevertheless, the results also show that, whereas 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 analyzed above.

SAXS data are often used to estimate R_g, 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. S10).

Assessing the influence of the prior on the parameters search

Because our strategy to determine self-consistent values for $\delta\rho$ and r_0 is based on the BME refinement of probability distributions, it is reasonable to ask how the results depend on the statistical prior used in the approach. In this context, there are two related questions that we address here. First, as we have also examined previously (65,67), is the question of how much the distribution of conformations after reweighting depends on the prior that is used. Second is the question of how much the $\delta\rho$ and r_0 parameters depend on the prior. The latter is important because the optimal parameters may in part compensate for imperfections in the prior.

To examine these questions we applied our protocol to three different ensembles of α-Synuclein. The first ensemble was generated using flexible-meccano, whereas the other two were previously generated by molecular dynamics simulations (66) using either the Amber a99SB-*disp* or the Amber ff03ws (a03ws) force field. The distributions of the R_g for the three priors are relatively different (Fig. 4 *a*), and consequently, the minima of the γ parameter indicate small differences in the best values of $\delta\rho$ and r_0 (Fig. 5). The reweighted distributions of R_g, however, appear very similar (Fig. 4 *b*). Notably, for each prior, we obtain essentially indistinguishable distributions of R_g whether we use the optimal



	🟦	🟧	🟩	🟥
$\delta\rho\,[e/nm^3]$	10.0	10.0	26.7	0.0
r_0 [Å]	1.558	1.394	1.394	7.722
χ^2_{red}	1.024	1.028	1.060	1.030
ϕ_{eff}	0.94	0.93	0.89	0.88

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 (*blue*) as well as three other choices of r_0 and $\delta\rho$ in the low-γ region (*orange*, *green*, and *red*). The insert shows the parameters used in each case and the results of the reweighting on the R_g distribution. To see this figure in color, go online.

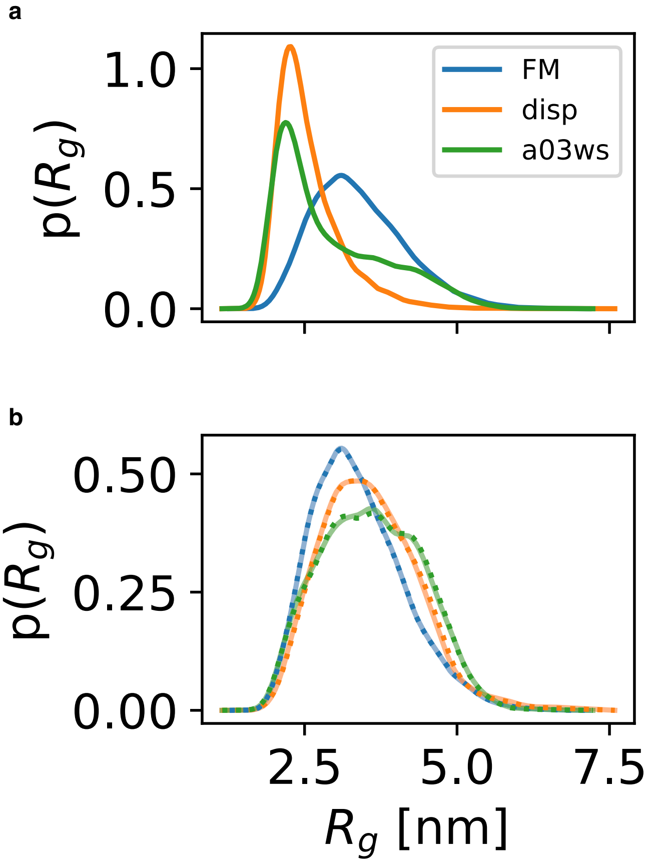


FIGURE 4 Effect of the prior distribution. (*a*) Distributions of R_g of α-Synuclein sampled with flexible-meccano (FM), a99SB-*disp* (disp), and a03ws. (*b*) Reweighted R_g distributions, either from the optimal (*lowest* γ) $\delta\rho$ and r_0 parameters for each ensemble (*solid lines*) or using the default values, we propose ($\delta\rho$= 3.34 e/nm^3 and $r_0 = 1.681$ Å; *dotted lines*). To see this figure in color, go online.

parameters (for each prior) from the grid search or the set of parameters that we proposed as default values (Fig. 4 *b*).

Returning to the original two questions, these results show that the prior may influence the optimal parameters resulting from the grid search, similar to our observations using synthetic SAXS data (Figs. S6 and S7). They also show, in line with previous observations (65,67), that even when starting from somewhat different priors, the posterior distributions tend to be substantially similar. Noteworthy, the results are robust to the choice of $\delta\rho$ and r_0, so that very similar results are obtained, even when using the global minimum from our analyses of Hst5, Sic1, Tau, and TIA1.

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 center of mass of the protein. This is a geometric quantity that is often used to study protein behavior and biophysics. Because the ensembles were constructed by fitting to the experimental SAXS data, the resulting averages and distributions of R_g represent the experimental system, but exactly because the hydration effects were included in the SAXS calculations, this means that these R_g-values only represent the protein.

Another approach to estimate $\langle R_g \rangle$ 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 (59), although other approaches exist (80–82). Because the SAXS data potentially contain 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 (17). 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 maximal 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 (83).

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, although the two might differ because of effects of the hydration layer. Thus, we performed a Guinier analysis of the SAXS data for the four proteins, progressively extending the upper limit of the q-range from 0.9 to 1.3 and plotting R_g vs. $q_{\max}R_g$ (83). We find that the Guinier fits can show substantial differences in the estimated $\langle R_g \rangle$ values, depending on the range used. Returning to the question of how the R_g-values estimated from the Guinier fit compare to the average R_g from the conformational ensembles with the lowest γ scores (*horizontal black line* in Fig. 6), we find that these are in a reasonable agreement (within 0.2 nm) with the values calculated from Guinier fits using $q_{\max}R_g = 1.3$. Looking across the four proteins, we do not find a unique $q_{\max}R_g$-value for which the Guinier fit gives rise to an average R_g that is similar to that obtained from the conformational ensembles.

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 on a forward model, such as Pepsi-SAXS (14), to calculate SAXS data from one or more conformations and optimize the structures or weights to improve agreement with experiments. Although these approaches are very powerful, they are subject to uncertainty due to the choice of unknown parameters in the forward model. In principle,

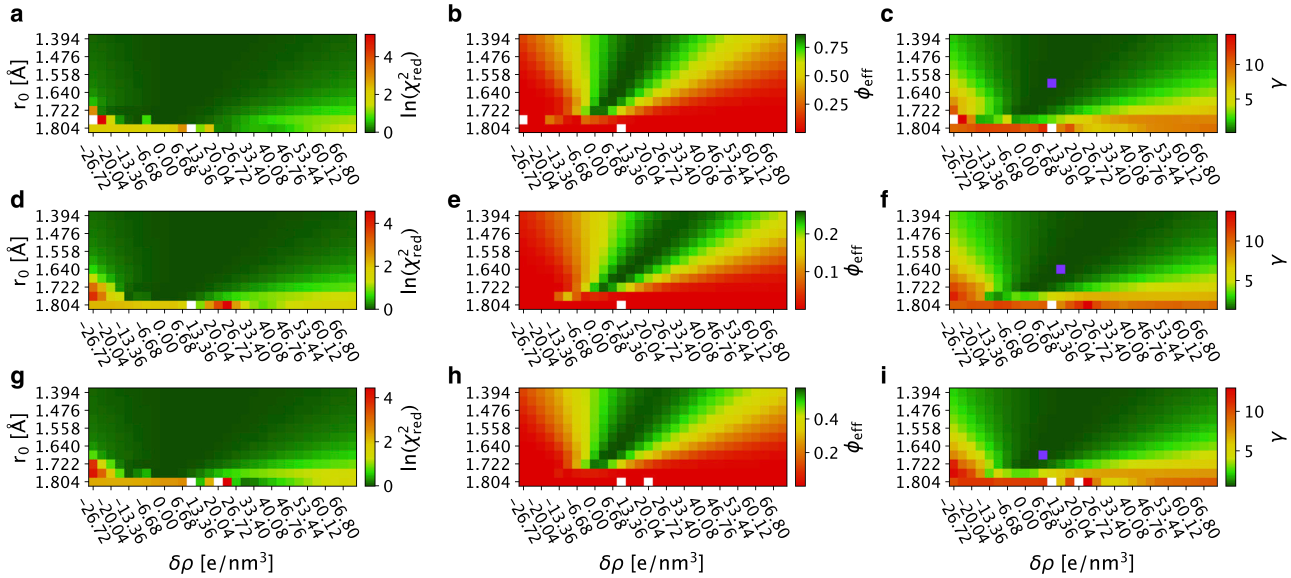


FIGURE 5 Reweighting α-Synuclein ensembles using SAXS data calculated using different values for the parameters that effect the contribution from the hydration layer and displaced solvent. The grids show the results from the iBME ensemble optimization with different combinations of $\delta\rho$ and r_0. The top row (*a–c*) shows the results from the flexible-meccano ensemble, the second row (*d–f*) shows the results using a99SB-disp as the prior, and the third row (*g–i*) shows the results from a03ws as the prior. For each ensemble we show in the first column (*a*, *d*, and *g*) $\ln(\chi^2_{\mathrm{red}})$, in the second column we show (*b*, *e*, and *h*) ϕ_{eff}, and in the third column (*c*, *f*, and *i*) we show $\gamma = \ln\left(\frac{\chi^2_{\mathrm{red}}}{\phi_{\mathrm{eff}}}\right)$. White spots correspond to ensembles in which the iBME reweighting failed. Purple spots in the third column correspond to the minima for γ. To see this figure in color, go online.

these parameters can be “integrated out” using Bayesian approaches (31,84), although this can become computationally prohibitive for SAXS calculations. Thus, the aim of our work is 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 *cst*) that are often necessary to estimate.

We have developed and tested iBME as an extension to 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. Although we have presented iBME here in the context of SAXS data, other types of data, such as NMR residual dipolar couplings, solvent paramagnetic relaxation enhancement effects, or circular dichroism spectra, 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 in which the ensembles are

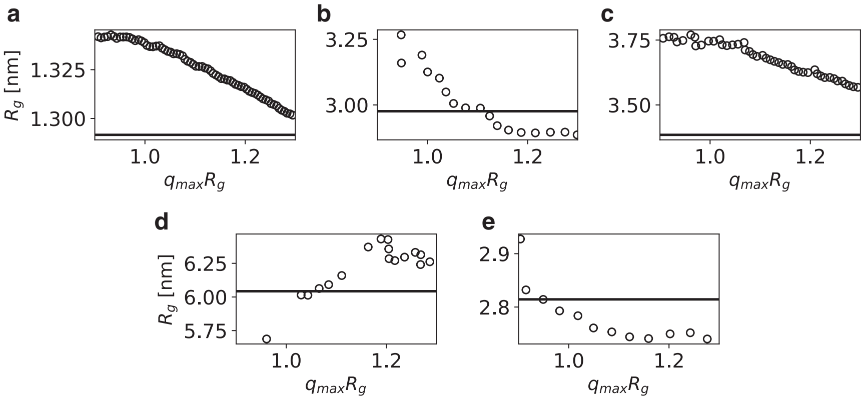


FIGURE 6 Estimating $\langle R_g \rangle$ from experimental SAXS profiles of (*a*) Hst5, (*b*) Sic1, (*c*) α-Synuclein, (*d*) Tau, and (*e*) TIA1 using Guinier fitting and ensemble refinement. We used the Guinier approximation to estimate R_g by fitting from the lowest measured value of q (in the case of Hst5 we ignored the first 10 points due to noise) to different values of q_{max}, reporting the results as R_g vs. $q_{\mathrm{max}}R_g$ (*black circles*). The horizontal black lines are the ensemble-averaged R_g calculated from the conformational ensembles (in the case of α-Synuclein, we used the flexible-meccano prior) with the chosen optimal r_0 and $\delta\rho$ parameters (Table 1).

reweighted using SAXS data calculated using different values for these parameters. Such an analysis is, in particular, important for large ensembles of flexible molecules because fitting these parameters to each conformation could lead to substantial overfitting. We also note that the calculations of SAXS intensities could potentially be improved further by being able to predict the features and contribution from the hydration layer for different sequences and conformations rather than relying on fitting parameters.

Combining these two aspects, the approach that we have described can be summarized as follows: 1) sampling a conformational ensemble; 2) calculating SAXS profiles from the conformers of the ensemble, keeping scale and background parameters fixed ($I(0) = 1$, $cst = 0$) and performing a grid scan for $\delta\rho$ and r_0; 3) for each value of $\delta\rho$ and r_0, optimizing the weights, $I(0)$ and cst using iBME and 4) examining the results by calculating χ^2_{red}, ϕ_{eff}, and γ, and selecting 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. In cases where the prior ensemble already exists or is fast to generate, the SAXS calculations can quickly become limiting in terms of computational efficiency. For these reasons we also propose default values for $\delta\rho$ and r_0 that we find to provide relatively accurate results for the four proteins that we examined. To test this further, we also used these default parameters to calculate SAXS intensities from different conformational ensembles of α-Synuclein and find that the resulting distributions of R_g are almost the same as if the parameters are optimized. We also note that the computational overhead of the grid scans could be drastically reduced by precomputing partial SAXS intensities once per grid and then adding the contributions from $\delta\rho$ and r_0. Although this procedure is already internally used by several methods to calculate SAXS data, options to output and process partial intensities for specific scattering angles are, at the moment, not easily accessible.

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. Although 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 both changing the r_0 and $\delta\rho$ parameters (Fig. S10) or the region used for Guinier fitting (Fig. 6) can change the R_g substantially, 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 because one needs to choose parameters in the SAXS calculations. Thus, we 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 (17) and suggest default values that can be used as a starting point.

SUPPORTING MATERIAL

Supporting material can be found online at https://doi.org/10.1016/j.bpj.2021.10.003.

AUTHOR CONTRIBUTIONS

F.P. and K.L.-L. designed research. F.P. performed research. F.P. and K.L.-L. analyzed data. F.P. and K.L.-L. wrote the manuscript.

ACKNOWLEDGMENTS

We acknowledge Sandro Bottaro for fruitful discussions and for implementing the iBME algorithm in the BME software, Simone Orioli for useful discussions and input about small-angle x-ray scattering 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 small-angle x-ray scattering data for, respectively, Tau and Sic1.

This research was funded by the Lundbeck Foundation BRAINSTRUC initiative in structural biology (R155-2015-2666 to K.L.L.)

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 Lett.* 589:2570–2577.
2. Grant, T. D. 2018. Ab initio electron density determination directly from solution scattering data. *Nat. Methods.* 15:191–193.
3. Prior, C., O. R. Davies, …, E. Pohl. 2020. Obtaining tertiary protein structures by the ab initio interpretation of small angle X-ray scattering data. *J. Chem. Theory Comput.* 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. *Curr. Opin. Struct. Biol.* 49:18–26.
6. Tria, G., H. D. T. Mertens, …, D. I. Svergun. 2015. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ.* 2:207–217.
7. Svergun, D. I., S. Richard, …, G. Zaccai. 1998. Protein hydration in solution: experimental observation by x-ray and neutron scattering. *Proc. Natl. Acad. Sci. USA.* 95:2267–2272.
8. Park, S., J. P. Bardhan, …, L. Makowski. 2009. Simulated x-ray scattering of protein solutions using explicit-solvent models. *J. Chem. Phys.* 130:134114.
9. Chen, P. C., and J. S. Hub. 2014. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. *Biophys. J.* 107:435–447.
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 Res.* 43:W225-30.
11. Köfinger, J., and G. Hummer. 2013. Atomic-resolution structural information from scattering experiments on macromolecules in solution. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 87:052712.
12. Grishaev, A., L. Guo, …, A. Bax. 2010. Improved fitting of solution x-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. *J. Am. Chem. Soc.* 132:15484–15486.

13. 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. *J. Appl. Cryst.* 28:768–773.

14. 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 Crystallogr. D Struct. Biol.* 73:449–464.

15. Schneidman-Duhovny, D., M. Hammel, and A. Sali. 2010. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* 38:W540–W544.

16. Schneidman-Duhovny, D., M. Hammel, …, A. Sali. 2013. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* 105:962–974.

17. Henriques, J., L. Arleth, …, M. Skepö. 2018. On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations. *J. Mol. Biol.* 430:2521–2539.

18. Best, R. B., W. Zheng, and J. Mittal. 2014. Balanced protein – water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* 10:5113–5124.

19. Piana, S., A. G. Donchev, …, D. E. Shaw. 2015. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B.* 119:5113–5123.

20. Henriques, J., C. Cragnell, and M. Skepö. 2015. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *J. Chem. Theory Comput.* 11:3420–3431.

21. Rauscher, S., V. Gapsys, …, H. Grubmüller. 2015. Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J. Chem. Theory Comput.* 11:5513–5524.

22. Palazzesi, F., M. K. Prakash, …, A. Barducci. 2015. Accuracy of current all-atom force-fields in modeling protein disordered states. *J. Chem. Theory Comput.* 11:2–7.

23. Fisher, C. K., A. Huang, and C. M. Stultz. 2010. Modeling intrinsically disordered proteins with Bayesian statistics. *J. Am. Chem. Soc.* 132:14919–14927.

24. Różycki, B., Y. C. Kim, and G. Hummer. 2011. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure.* 19:109–116.

25. Beauchamp, K. A., V. S. Pande, and R. Das. 2014. Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys. J.* 106:1381–1390.

26. Hummer, G., and J. Köfinger. 2015. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* 143:243150.

27. Bonomi, M., C. Camilloni, …, M. Vendruscolo. 2016. Metainference: a Bayesian inference method for heterogeneous systems. *Sci. Adv.* 2:e1501177.

28. Shevchuk, R., and J. S. Hub. 2017. Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. *PLoS Comput. Biol.* 13:e1005800.

29. Potrzebowski, W., J. Trewhella, and I. Andre. 2018. Bayesian inference of protein conformational ensembles from limited structural data. *PLoS Comput. Biol.* 14:e1006641.

30. Lincoff, J., M. Haghighatlari, …, T. Head-Gordon. 2020. Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Commun. Chem.* 3:74.

31. Spill, Y. G., Y. Karami, …, M. Nilges. 2021. Automatic Bayesian weighting for SAXS data. *Front. Mol. Biosci.* 8:671011.

32. Pitera, J. W., and J. D. Chodera. 2012. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* 8:3445–3451.

33. Boomsma, W., J. Ferkinghoff-Borg, and K. Lindorff-Larsen. 2014. Combining experiments and simulations using the maximum entropy principle. *PLOS Comput. Biol.* 10:e1003406.

34. Cesari, A., S. Reißer, and G. Bussi. 2018. Using the maximum entropy principle to combine simulations and solution experiments. *Computation (Basel).* 6:15.

35. Hermann, M. R., and J. S. Hub. 2019. SAXS-restrained ensemble simulations of intrinsically disordered proteins with commitment to the principle of maximum entropy. *J. Chem. Theory Comput.* 15:5103–5115.

36. 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.

37. Orioli, S., A. H. Larsen, …, K. Lindorff-Larsen. 2020. Chapter Three - How to learn from inconsistencies: integrating molecular simulations with experimental data. *In* Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly. B. Strodel and B. Barz, eds. Academic Press, pp. 123–176.

38. Bernadó, P., and D. I. Svergun. 2012. Structural analysis of intrinsically disordered proteins by small-angle x-ray scattering. *Mol. Biosyst.* 8:151–167.

39. Pelikan, M., G. L. Hura, and M. Hammel. 2009. Structure and flexibility within proteins as identified through small angle x-ray scattering. *Gen. Physiol. Biophys.* 28:174–189.

40. Krzeminski, M., J. A. Marsh, …, J. D. Forman-Kay. 2013. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics.* 29:398–399.

41. Yang, S., L. Blachowicz, …, B. Roux. 2010. Multidomain assembled states of Hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci. U.S.A.* 107:15757–15762.

42. Brüschweiler, R., and D. Case. 1994. Adding harmonic motion to the Karplus relation for spin-spin coupling. *J. Am. Chem. Soc.* 116:11199–11200.

43. Lindorff-Larsen, K., R. B. Best, and M. Vendruscolo. 2005. Interpreting dynamically-averaged scalar couplings in proteins. *J. Biomol. NMR.* 32:273–280.

44. Louhivuori, M., R. Otten, …, A. Annila. 2006. Conformational fluctuations affect protein alignment in dilute liquid crystal media. *J. Am. Chem. Soc.* 128:4371–4376.

45. Salvatella, X., B. Richter, and M. Vendruscolo. 2008. Influence of the fluctuations of the alignment tensor on the analysis of the structure and dynamics of proteins using residual dipolar couplings. *J. Biomol. NMR.* 40:71–81.

46. Vitkup, D., D. Ringe, …, G. A. Petsko. 2002. Why protein R-factors are so large: a self-consistent analysis. *Proteins.* 46:345–354.

47. Moore, P. B. 2014. The effects of thermal disorder on the solution-scattering profiles of macromolecules. *Biophys. J.* 106:1489–1496.

48. Meisburger, S. P., W. C. Thomas, …, N. Ando. 2017. X-ray scattering studies of protein structural dynamics. *Chem. Rev.* 117:7615–7672.

49. Xu, D., S. P. Meisburger, and N. Ando. 2021. Correlated motions in structural biology. *Biochemistry.* 60:2331–2340.

50. Ozenne, V., F. Bauer, …, 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.

51. Estaña, A., N. Sibille, …, P. Bernadó. 2019. Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database. *Structure.* 27:381–391.e2.

52. Jensen, M. R., P. R. Markwick, …, M. Blackledge. 2009. Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure.* 17:1169–1185.

53. Bernadó, P., L. Blanchard, …, M. Blackledge. 2005. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proc. Natl. Acad. Sci. USA.* 102:17002–17007.

54. Wells, M., H. Tidow, …, A. R. Fersht. 2008. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl. Acad. Sci. USA.* 105:5762–5767.

55. Mukrasch, M. D., P. Markwick, …, M. Blackledge. 2007. Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *J. Am. Chem. Soc.* 129:5235–5243.

56. Rotkiewicz, P., and J. Skolnick. 2008. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* 29:1460–1465.

57. McGibbon, R. T., K. A. Beauchamp, …, V. S. Pande. 2015. MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* 109:1528–1532.

58. Fraser, R. D. B., T. P. MacRae, and E. Suzuki. 1978. An improved method for calculating the contribution of solvent to the x-ray diffraction pattern of biological molecules. *J. Appl. Cryst.* 11:693–694.

59. Guinier, A. 1939. La diffraction des rayons X aux très petits angles: application à l'étude de phénomènes ultramicroscopiques. *Ann. Phys. (Paris).* 11:161–237.

60. Pedregosa, F., G. Varoquaux, …, E. Duchesnay. 2011. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12:2825–2830.

61. Jephthah, S., L. Staby, …, M. Skepö. 2019. Temperature dependence of intrinsically disordered proteins in simulations: what are we missing? *J. Chem. Theory Comput.* 15:2672–2683.

62. Gomes, G.-N. W., M. Krzeminski, …, C. C. Gradinaru. 2020. Integrating multiple experimental data to determine conformational ensembles of an intrinsically disordered protein. *bioRxiv* https://doi.org/10.1101/2020.02.05.935890.

63. Mylonas, E., A. Hascher, …, D. I. Svergun. 2008. Domain conformation of tau protein studied by solution small-angle x-ray scattering. *Biochemistry.* 47:10345–10353.

64. Sonntag, M., P. K. A. Jagtap, …, M. Sattler. 2017. Segmental, domain-selective perdeuteration and small-angle neutron scattering for structural analysis of multi-domain proteins. *Angew. Chem. Int.Engl.* 56:9322–9325.

65. Ahmed, M. C., L. K. Skaanning, …, K. Lindorff-Larsen. 2021. Refinement of α-synuclein ensembles against SAXS data: comparison of force fields and methods. *Front. Mol. Biosci.* 8:654333.

66. Robustelli, P., S. Piana, and D. E. Shaw. 2018. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA.* 115:E4758–E4766.

67. Larsen, A. H., Y. Wang, …, 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 Comput. Biol.* 16:e1007870.

68. Monticelli, L., S. K. Kandasamy, …, S.-J. Marrink. 2008. The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* 4:819–834.

69. Virtanen, J. J., L. Makowski, …, K. F. Freed. 2010. Modeling the hydration layer around proteins: HyPred. *Biophys. J.* 99:1611–1619.

70. Virtanen, J. J., L. Makowski, …, K. F. Freed. 2011. Modeling the hydration layer around proteins: applications to small- and wide-angle x-ray scattering. *Biophys. J.* 101:2061–2069.

71. Persson, F., P. Söderhjelm, and B. Halle. 2018. The geometry of protein hydration. *J. Chem. Phys.* 148:215101.

72. van Gunsteren, W. F., X. Daura, …, L. J. Smith. 2018. Validation of molecular simulation: an overview of issues. *Angew. Chem. Int.Engl.* 57:884–902.

73. Rambo, R. P., and J. A. Tainer. 2013. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature.* 496:477–481.

74. 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. *J. Appl. Cryst.* 54:1281–1289.

75. Hansen, S. 2000. Bayesian estimation of hyperparameters for indirect fourier transformation in small-angle scattering. *J. Appl. Cryst.* 33:1415–1421.

76. 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.

77. Qian, H. 2001. Relative entropy: free energy associated with equilibrium fluctuations and nonequilibrium deviations. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 63:042103.

78. Zamyatnin, A. A. 1972. Protein volume in solution. *Prog. Biophys. Mol. Biol.* 24:107–123.

79. Roche, J., and C. A. Royer. 2018. Lessons from pressure denaturation of proteins. *J. R. Soc. Interface.* 15:20180244.

80. Vestergaard, B., and S. Hansen. 2006. Application of Bayesian analysis to indirect Fourier transformation in small-angle scattering. *J. Appl. Cryst.* 39:797–804.

81. Riback, J. A., M. A. Bowman, …, T. R. Sosnick. 2017. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science.* 358:238–241.

82. Zheng, W., and R. B. Best. 2018. An extended Guinier analysis for intrinsically disordered proteins. *J. Mol. Biol.* 430:2540–2553.

83. Borgia, A., W. Zheng, …, B. Schuler. 2016. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J. Am. Chem. Soc.* 138:11714–11726.

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

Biophysical Journal, Volume *120*

Supplemental information

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

Francesco Pesce and Kresten Lindorff-Larsen

Supplementary material: 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

1 Validation of iBME with synthetic SAXS data

We used the Hst5 ensembles generated by Flexible-Meccano to create synthetic data to test and validate iBME. We calculated SAXS profiles for each structure in the ensemble using Pepsi-SAXS, using $\delta\rho = 13.36\ e/nm^3$ and $r_0 = 1.722$ Å as parameters to describe the hydration layer and displaced solvent volumes. These SAXS profiles were linearly averaged to give rise to the target SAXS data, assigning the error associated with the j'th data point as $\sigma_j = \frac{0.5 I_j}{100}\exp(q_j)$

We then generated five sets of non-uniform weights for the ensemble, to define five different prior distributions. Specifically, we generated weights for each conformer, i, according to:

$$w_i' = \exp\{1 - (20 + 10a) * \exp[(0.4b + 1.2) - R_g(i))^4]\} \tag{1}$$

with a and b being random numbers between 0 and 1, and with the final weights (w_i) obtained by normalizing w_i'. These weights lead to the SAXS data in Fig. S1) and R_g distributions shown in Fig. S2.

First we use the standard BME approach with $\theta = 100$ to optimize each of the five priors against the (synthetic) experimental data by minimizing the functional

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

as described in the main text and with χ^2_{red} defined as

$$\chi^2_{\mathrm{red}}(\omega_1 \cdots \omega_n) = \frac{1}{m}\sum_i^m \frac{(\sum_j^n \omega_j F(x_j) - F_i^{EXP})^2}{\sigma_i^2} \tag{3}$$

After doing so we keep the resulting weights (ω_j) as reference.

Standard BME assumes that the experimental observable (F^{EXP}) are on the same scale as those obtained by applying the forward model to the computational ensemble ($F(x_j)$). iBME is an approach that deals with cases where the two are on a different scale, such as for example SAXS data where I^{EXP} and the calculated $I(x_j)$ may differ by a linear transformation $I'(x_j) = \text{scale} \cdot I(x_j) + \text{offset}$.

To generate synthetic data representing this situation, we thus changed each of the input SAXS curves (for each of the structures) by multiplying by a random number between 0 and 5 to change the scale, and subsequently adding a random number between 0 and 1 to change the offset (the same scale and offset was used for each structure in the ensemble). The average curves are shown in Fig. S3.

We then applied iBME (as described in the main text) to these priors, targeting the unmodified synthetic data (blue line in Figs. S1 and S3). The same θ as with standard BME (100) was used. The successful outcome of iBME is demonstrated by comparing the χ^2 both before and after reweighting, ϕ_{eff} and and the weights obtained using BME on the unmodified SAXS data (Fig. S4).

2 Validation of the grid scan with synthetic SAXS data

We use the same synthetic experimental SAXS data used to test iBME above (i.e. to fit the scale and offset) to assess the ability of the grid scan procedure to recover the $\delta\rho$ and r_0 used to generate the synthetic experimental SAXS profile.

We first used uniform weights (same as used to generate the synthetic experimental SAXS profile) as the prior for the iBME optimization. Even when adding noise to the (synthetic) data, the grid search recovers the correct values used to generate the synthetic data ($\delta\rho = 13.36e/nm^3$ and $r_0 = 1.722$Å) at the minimum of γ (determined by a $\chi^2_{red} \approx 0$ and $\phi_{eff} \approx 1$) (Fig. S6).

We also repeated the grid scan using ‘Prior 1’ (Figs. S2–S3) discussed above as the prior. In this way, we represent the case where the experimental data are generated by a different distribution than the prior. In this case we find the minimum of γ in a grid point adjacent to that used to generate the synthetic experimental SAXS profile ($\delta\rho = 10.02\ e/nm^3$ and $r_0 = 1.763$Å) (Fig. S7). Thus, while we do not recover exactly the same values, they are very close to those used to generate the data.

3 Additional figures and tables

Table S1: Search ranges for fitting parameters

	CRYSOL	FoXS	Pepsi-SAXS
r_0 [Å]	1.55 - 1.68	1.40 - 1.80	1.56 - 1.72
$\delta\rho$ [$e\ nm^{-3}$]	0 - 70.0	-27.0 - 54.0	0 - 33.4

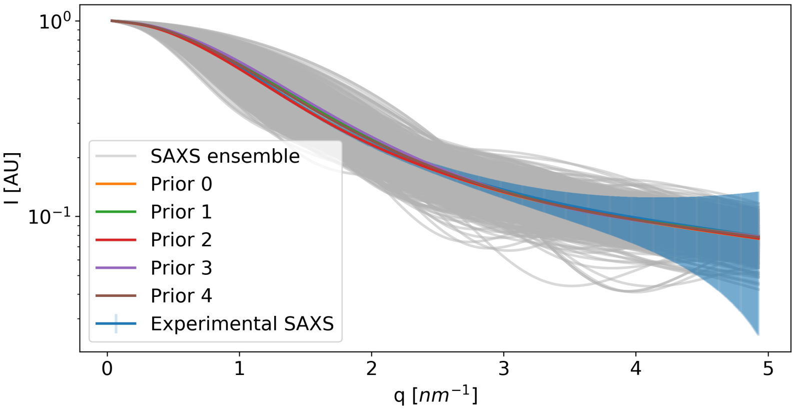


Figure S1: Synthetic SAXS data used to validate the iBME protocol. Thin grey lines show SAXS profiles for each of the structures in the Hst5 ensemble. The uniform average of these curves gives rise to the blue line, which we here term the (synthetic) ‘experimental’ data. This is the target for the optimization. The non-uniform weights give rise to five other average SAXS curves, that are the starting point for optimization.

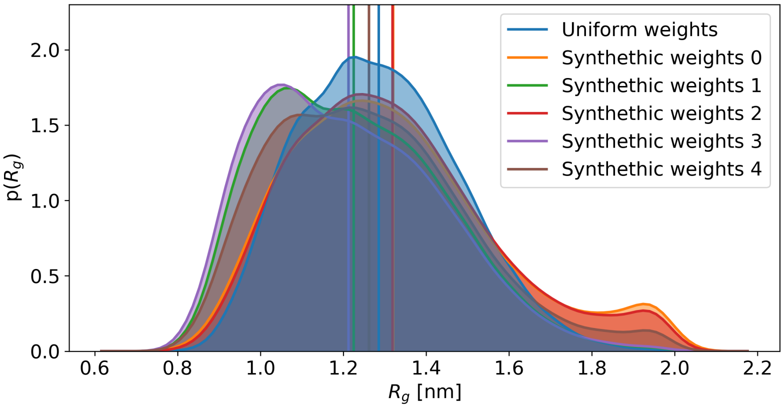


Figure S2: R_g distribution from the Flexible-meccano ensemble of Hst5 (with uniform weights), as well as five ensembles with different sets of non-uniform weights.

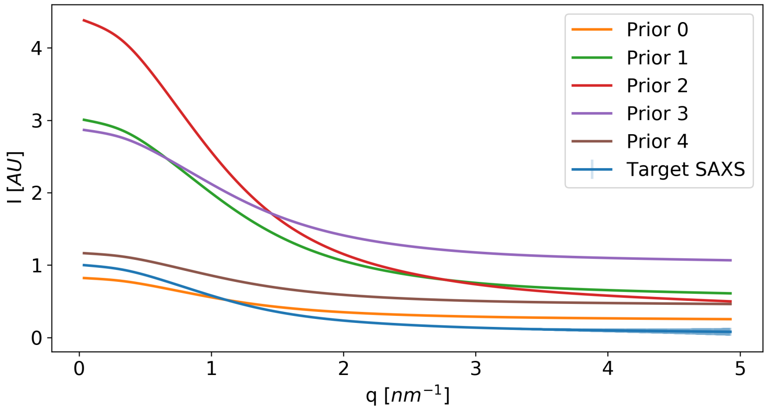


Figure S3: Same SAXS profiles as in Fig. S1, but after perturbations with a scale and offset.

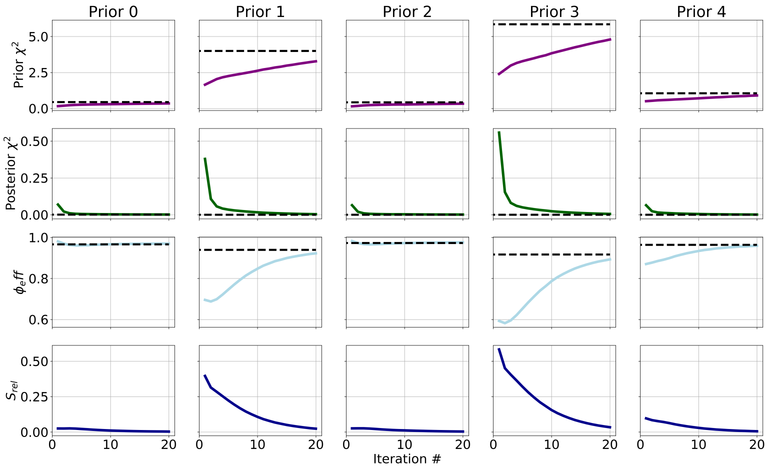


Figure S4: Evolution of observables along the iterations of the iBME. Dotted black lines represent the target values obtained from the standard BME using the un-scaled and shifted data. The relative entropy S_{rel} is computed between the weights at each iteration of iBME and those obtained from standard BME.

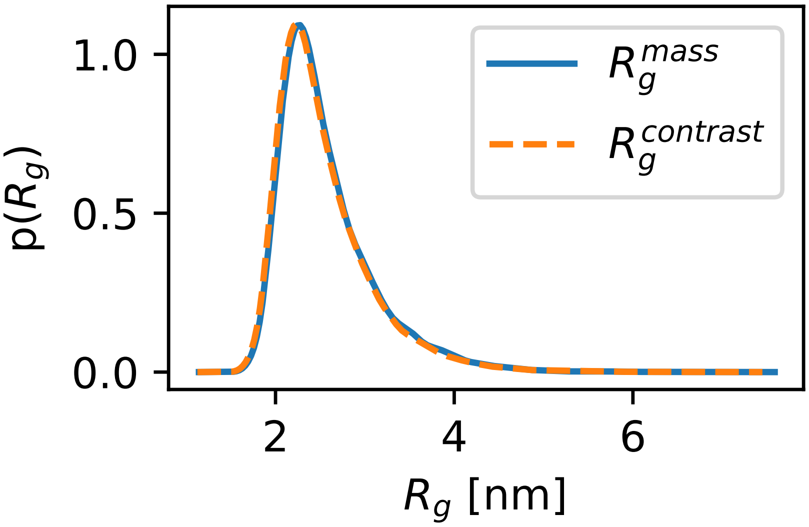


Figure S5: Distribution of the mass-weighted R_g and contrast-weighted R_g calculated as described in the Methods of the main text for the a99SB-disp ensemble of α-Synuclein.

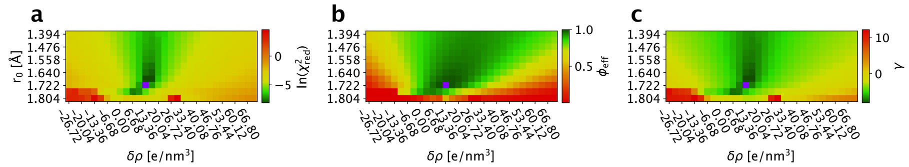


Figure S6: Grid scan optimizing a synthetic experimental SAXS profile with iBME. In this case we used as prior the same distribution as that used to generate the synthetic data. The minima in χ^2_{red} (a), ϕ_{eff} (b) and γ are shown in purple.

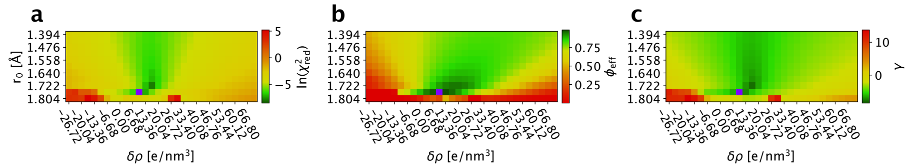


Figure S7: Grid scan optimizing a synthetic experimental SAXS profile with iBME. As prior for iBME we use 'Prior 1' (Figs. S2 and S3). Minima in χ^2_{red} (a), ϕ_{eff} (b) and γ are shown in purple.

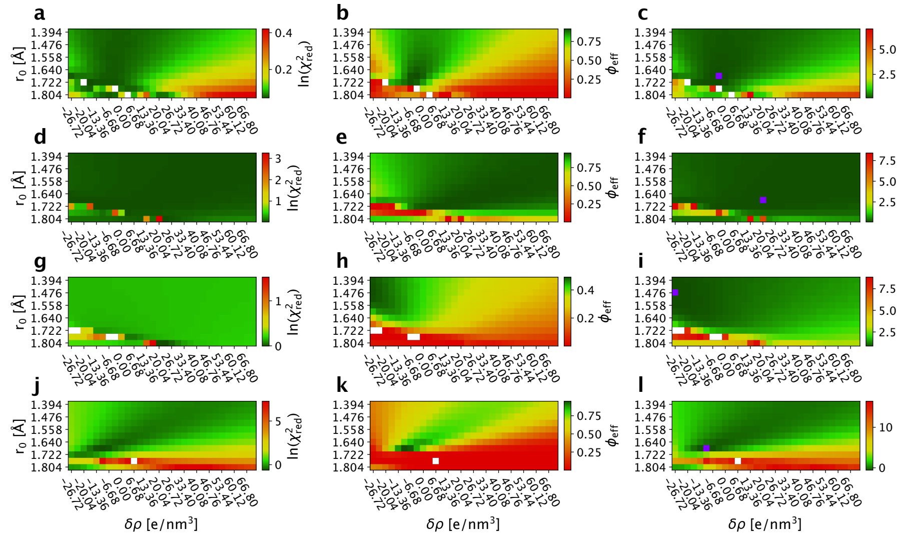


Figure S8: Reweighting ensembles using SAXS data calculated with FoXS using different values for the parameters that effect the contribution from for the hydration layer and displaced solvent. The grids show the results from the iBME ensemble optimization with different combinations of $\delta\rho$ and r_0. The top row (a–c) shows Hst5, the second row (d–f) shows Sic1, the third row (g–i) shows Tau, and the last row (j–l) shows results for TIA1. For each protein we show in the first column (a, d, g, j) $\ln\left(\chi^2_{\text{red}}\right)$, the second column (b, e, h, k) ϕ_{eff}, and third column (c, f, i, l) $\gamma = \ln\left(\frac{\chi^2_{\text{red}}}{\phi_{\text{eff}}}\right)$. White spots correspond to ensembles where the iBME reweighting failed. The purple spots in the third column correspond to the minima for γ.

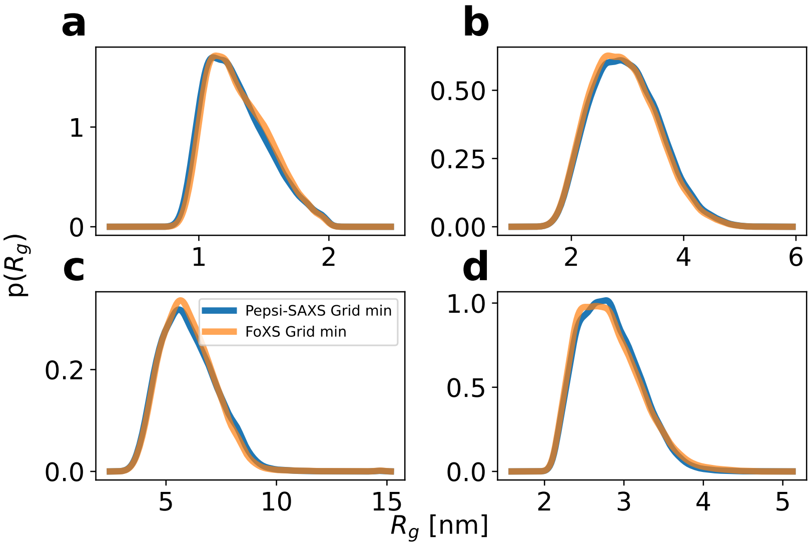


Figure S9: Reweighted R_g distributions for (a) Hst5, (b) Sic1, (c) Tau and (d) TIA-1 from the γ minima obtained with either Pepsi-SAXS or FoXS-based grid scans.

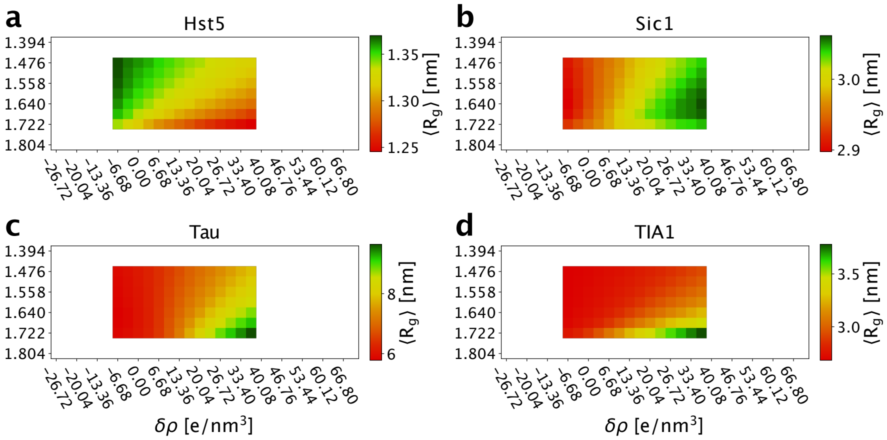


Figure S10: Reweighted average values of R_g on the part of the grids that gave reasonable fits for (a) Hst5, (b) Sic1, (c) Tau and (d) TIA-1.