CHAPTER TWO

Structure and ensemble refinement against SAXS data: Combining MD simulations with Bayesian inference or with the maximum entropy principle

Leonie Chatzimagas and Jochen S. Hub*
Theoretical Physics and Center for Biophysics, Saarland University, Saarbrücken, Germany
*Corresponding author: e-mail address: jochen.hub@uni-saarland.de

Contents

Methods in Enzymology, Volume 678
ISSN 0076-6879
https://doi.org/10.1016/bs.mie.2022.09.014

Abstract

Small-angle X-ray scattering (SAXS) is a powerful method for tracking conformational transitions of proteins or soft-matter complexes in solution. However, the interpretation of the experimental data is challenged by the low spatial resolution and the low information content of the data, which lead to a high risk of overinterpreting the data. Here, we illustrate how SAXS data can be integrated into all-atom molecular dynamics (MD) simulation to derive atomic structures or heterogeneous ensembles that are compatible with the data. Besides providing atomistic insight, the MD simulation adds physicochemical information, as encoded in the MD force fields, which greatly reduces the risk of overinterpretation. We present an introduction into the theory of SAXS-driven MD simulations as implemented in GROMACS-SWAXS, a modified version of the GROMACS simulation software. We discuss SAXS-driven parallel-replica ensemble refinement with commitment to the maximum entropy principle as well as a Bayesian formulation of SAXS-driven structure refinement. Practical considerations for running and interpreting the simulations are presented. The methods are freely available via GitLab at https://gitlab.com/cbjh/gromacs-swaxs.

1. Introduction

Understanding the function of biomolecules requires understanding of their conformational dynamics. An increasingly popular method for tracking conformational transition of biomolecules is small-angle X-ray scattering (SAXS), which provides structural information that is not accessible by other techniques. Unlike NMR spectroscopy, which probes local distances and angles of smaller biomolecules, SAXS provides information on the overall shape and is applicable to both small and large biomolecules. Unlike crystallography or cryoelectron microscopy, SAXS probes molecules at ambient temperatures in solution, enabling experiments that track conformational transition after application of external stimuli. The accuracy of SAXS data has greatly improved over recent years thanks to sample delivery coupled with size exclusion chromatography (SEC-SAXS), thereby reducing sample aggregation artifacts, single-photon counting detractors, and standards for sample preparation (Berthaud, Manzi, Pérez, & Mangenot, 2012; Jeffries et al., 2016). Software and algorithms for data analysis and for SAXS-based structural modeling has greatly developed (Gräwert & Svergun, 2020). These properties and developments establish SAXS as an indispensable tool for integrative structural biology (Brosey & Tainer, 2019; Rout & Sali, 2019).

The interpretation of the SAXS data is challenged by the low information content of the experimental signals. Because the biomolecules are

randomly oriented during solution scattering, SAXS curves $I(q)$ represent an orientational average and, consequently, are only a one-dimensional smooth function of momentum transfer $q = 4\pi \sin(\theta)/\lambda$. Here, λ is the X-ray wavelength and 2θ is the scattering angle. The number of data points in $I(q)$ that provide independent structural information is estimated by the number of Shannon–Nyquist channels (Moore, 1980; Rambo & Tainer, 2013)

$$N_{\text{Shan}} = (q_{\max} - q_{\min})D/\pi, \tag{1}$$

where $q_{\max}$ and $q_{\min}$ denote the maximum and minimum momentum transfer in $I(q)$, respectively, and D is the maximum diameter of the solute. N_{Shan} is in the range of 5–30 for many SAXS experiments. For comparison, even a small protein with 100 residues contains approximately 200 flexible backbone angles, demonstrating that SAXS data is by far insufficient for defining all degrees of freedom of a biomolecule. Consequently, structure refinement against SAXS data is highly ambiguous, that is, many different structures fit the data equally well. Challenges due to the low information content are enhanced by the presence of heterogeneous ensembles and by uncertainties in the experimental and predicted SAXS curves. The low information content of the data together with larger number of degrees of freedom leads to a significant risk of overinterpretation upon fitting structural models against experimental data (Hub, 2018; Putnam, Hammel, Hura, & Tainer, 2007).

To mitigate the risk of overinterpretation during structure refinement, two main strategies have been devised. First, nearly all degrees of freedom of the biomolecule have been constrained, leading to methods such as rigid-body or normal mode refinement (Gorba, Miyashita, & Tama, 2008; Pelikan, Hura, & Hammel, 2009; Petoukhov & Svergun, 2005). Such methods use only simple energy terms to discriminate plausible from prohibited conformations, such as volume exclusion terms between protein domains or a multidimensional harmonic potential along normal modes. Second, physicochemical information has been added to the SAXS data with the help of atomistic molecular dynamics (MD) simulations (Björling, Niebling, Marcellini, van der Spoel, & Westenhoff, 2015; Chen & Hub, 2015; Kimanius, Pettersson, Schluckebier, Lindahl, & Andersson, 2015; Paissoni, Jussupow, & Camilloni, 2020). During MD simulations, all degrees of freedom are kept flexible, but the force field restrains the biomolecule to realistic conformations with acceptable free energy. Simulations with a

coupling to experimental SAXS curves have been coined SAXS-driven or SAXS-guided MD simulations.

Several previous studies described SAXS-driven simulations, in which the MD simulations were coupled to experimental SAXS data with an energetic restraint. Björling et al. (2015) presented a method with a focus on the interpretation of SAXS curve *differences* between two conformational states. This study neglected effects from the hydration layer on the SAXS curve, rationalized by the fact that hydration layer effects may approximately cancel upon taking SAXS curve differences. Our group developed SAXS-driven MD focused on absolute SAXS curves and implemented the method into an extension of the GROMACS software, coined GROMACS-SWAXS (https://gitlab.com/cbjh/gromacs-swaxs). The implementation uses explicit-solvent SAXS curve calculations using atomistic representations for the hydration layer and excluded solvent (Chen & Hub, 2014, 2015). Kimanius et al. (2015) suggested refining protein structures against SAXS data using metadynamics. However, since the SAXS curve calculations neglected the buffer subtraction, the implementation was not yet ready for refinement against experimental data. Paissoni et al. (2020) provided a method for SAXS-driven simulations implemented in PLUMED, which was primarily motivated with the aim of accelerating SAXS-driven simulations. This method maps the atomistic model to a coarse-grained representation for computing a SAXS curve. The method neglects effects from the hydration layer, and the coarse-grained approximation is limited to smaller scattering angles (Bernetti & Bussi, 2021). However, as discussed in this chapter and previously, the SAXS-driven MD simulations implemented in GROMACS-SWAXS are subject to only a small computational overhead between 5% and 20%, suggesting that SAXS-driven MD simulations based on all-atom SAXS predictions are likewise computationally feasible. Hsu, Leshchev, Kosheleva, Kohlstedt, and Chen (2020) presented SAXS-driven MD simulations utilizing the Debye equation, similar to Björling et al. (2015); however, these authors included effects from the hydration layer by increasing atomic form factors of solvent-exposed atoms, similar to the FoXS method (Schneidman-Duhovny, Hammel, & Sali, 2010).

This chapter describes the refinement of protein and soft-matter complexes by coupling all-atom MD simulations to experimental SAXS data, as implemented in GROMACS-SWAXS. We first describe SAXS-driven MD simulations coupled with a harmonic restraint to the data (Chen & Hub, 2015). The method was extended for simultaneous refinement against SAXS and several small-angle neutron scattering (SANS) data sets collected

at different D_2O concentrations (Chen et al., 2019). We discuss how SAXS-guided structure refinement is embedded into rigorous probability theory, as the obtained MD ensemble may be interpreted as the posterior distribution of a Bayesian inference problem (Shevchuk & Hub, 2017). Such viewpoint enables statements of confidence intervals, which are still underdeveloped in SAXS-guided modeling. Finally, we describe the refinement of heterogeneous ensembles against SAXS data using a minimal bias, that is, with commitment to Jaynes' principle of maximum entropy (Hermann & Hub, 2019; Ivanović, Hermann, Wójcik, Pérez, & Hub, 2020).

The methods described here are freely available via an extension of the GROMACS simulation software (Abraham et al., 2015) developed at https://gitlab.com/cbjh/gromacs-swaxs. We present the theoretical basis of SAXS-driven MD simulations. We discuss conceptual considerations as well as practical guidelines for running and interpreting the calculations. Additional tutorials and documentation are available at https://cbjh.gitlab.io/gromacs-swaxs-docs. Although we here refer mostly to SAXS, the methods described here are likewise applicable for refinement against small-angle neutron scattering (SANS) data or for refinement against combined SAXS/SANS data.

SAXS-driven MD simulations require a forward model for predicting SAXS curves from given MD simulations. The methods implemented in GROMACS-SWAXS utilize explicit-solvent calculations, thereby taking explicit representations of the hydration layer and excluded solvent into account (Chen & Hub, 2014; Merzel & Smith, 2002; Park, Bardhan, Roux, & Makowski, 2009). For more details on explicit-solvent SAXS predictions, we refer to chapter "Predicting solution scattering patterns with explicit-solvent molecular simulations" by Chatzimagas and Hub in Part A of this monograph.

2. SAXS-driven molecular dynamics simulations

2.1 Experiment-supported energetic bias

Similar to refinement against other experimental data (Jack & Levitt, 1978), SAXS-guided refinement is implemented by augmenting the MD force field energy $V_{FF}(R)$ with an experiment-derived energy $E_{exp}(R, D)$ that drives the simulation into conformations R that are compatible with the data D:

$$E_{hybrid} = V_{FF}(R) + E_{exp}(R,D) \tag{2}$$

Here, the data is given by the experimental SAXS intensities $I_{exp}(q_i)$ and errors $\sigma(q_i)$, where $i = 1,\ldots,N_{exp}$ and N_{exp} is the number of experimental

data points. The MD simulation is coupled to the data using a harmonic restraint on the SAXS intensities following

$$E_{\mathrm{exp}}(R(t), D) = \frac{1}{2} f_c \, k_B T \sum_{i=1}^{N_{\mathrm{exp}}} \frac{\left[I_c(q_i; R(t)) - (f I_{\mathrm{exp}}(q_i) + c)\right]^2}{\left[f \sigma(q_i)\right]^2}. \qquad (3)$$

Here, I_c is the SAXS curve calculated on-the-fly from the MD simulation and f_c is an overall force constant, whereas k_B and T denote the Boltzmann constant and the temperature. The factor 1/2 is introduced to enable a Bayesian interpretation of the refined ensemble in the special case $f_c = 1$ as described below. The symbols f and c denote fitting parameter that adjust the overall scale and an offset of the experimental curve relative to the calculated curve; f and c are adjusted throughout the simulation by minimizing E_{exp}. Hence, only those differences between I_c and I_{exp} that cannot be explained by f and c contribute to E_{exp} and, thereby, drive the simulation. The offset c absorbs uncertainties due to the buffer subtraction and may be omitted when coupling to high-precision SAXS data.

As an example, Fig. 1 presents structure refinement of the nuclear exportin CRM1, which transports protein cargos across the nuclear pore (Chen & Hub, 2015). Free simulations of CRM1 are highly force field dependent, evident from the fact that simulations with the Amber99sb or with the Charmm22* force fields lead to overly open or overly closed conformations, respectively (Fig. 1A, B, D top). Simulations with each force field exhibit poor agreement with experimental SAXS data (Fig. 1C, black and solid yellow or green curves). Upon restraining the simulations to SAXS data, excellent agreement with the data is obtained (Fig. 1C, dashed yellow or green curves), and both two force fields lead to a consensus partly open conformation (Fig. 1D bottom). Hence, the SAXS data overrules imperfections in the two force fields. Fig. 1C shows large residuals at low q between the experimental and the computed SAXS curves. These large residuals are a consequence of very small experimental errors, while the overall error is likely dominated by systematic errors. If systematic errors would be ignored, such large residuals would lead to excessively large SAXS-derived forces, hence asking for methods for modeling systematic errors as described below (see Sections 2.3 and 3.2).

2.2 Increasing the computational efficiency by smoothing and re-binning the experimental curve

Owing to the large number of pixels on modern X-ray detectors, $I_{\mathrm{exp}}(q_i)$ is heavily oversampled. $I_{\mathrm{exp}}(q_i)$ contains typically 800–2500 noisy estimates of

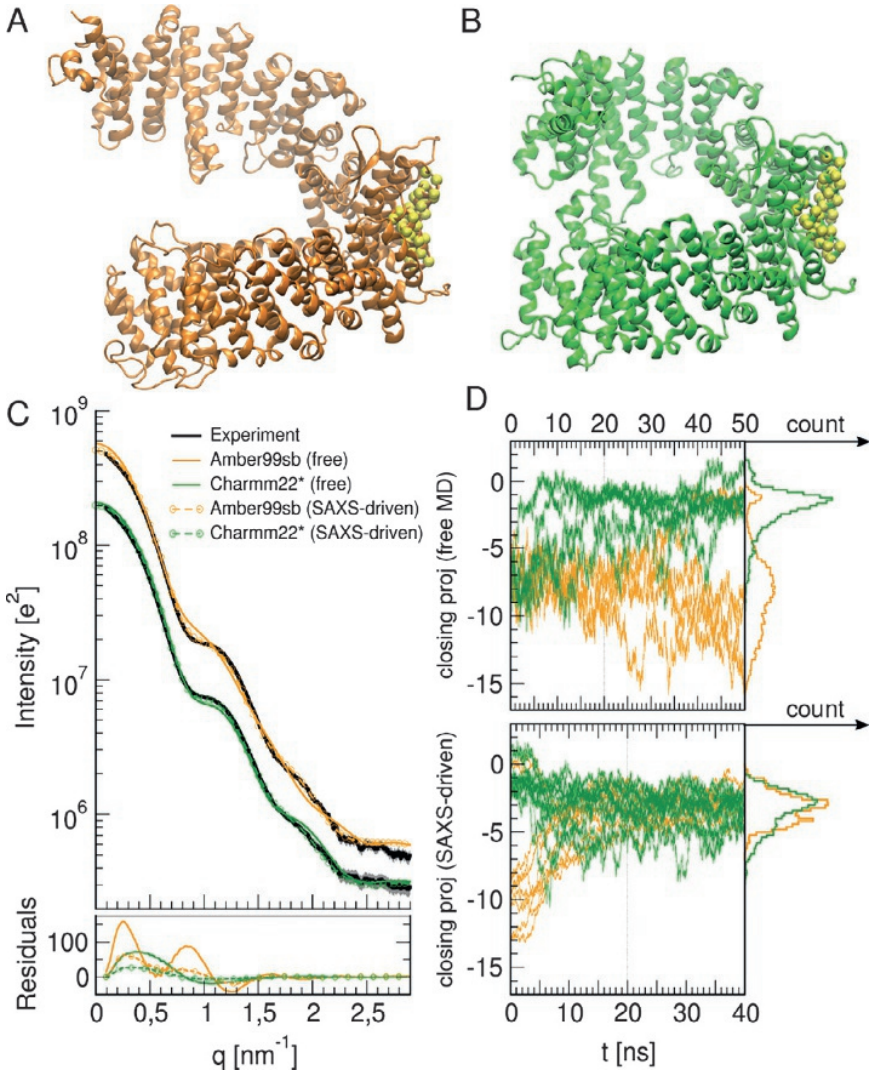

Fig. 1 Structure refinement of the nuclear exportin CRM1 against SAXS data. Free MD simulations with the Amber99sb (*yellow*) or Charmm22* (*green*) force fields lead to different conformations (A/B; D, *top*), which both disagree with experimental SAXS data ($\chi^2 = 1397$ and $\chi^2 = 429$ for Amber99sb and Charmm22*, respectively) (C, see legend). During SAXS-driven simulations, the different force fields lead to similar conformations (D, *bottom*), in excellent agreement with the data ($\chi^2 = 263$ and $\chi^2 = 57$ for Amber99sb and Charmm22*, respectively). *Adapted and reused with permission from Chen, P.-c., & Hub, J. S. (2015). Interpretation of solution X-ray scattering by explicit-solvent molecular dynamics. Biophysical Journal, 108, 2573–2584.*

the true underlying smooth SAXS curve, which contains only few independent data points ($N_{\mathrm{exp}} \gg N_{\mathrm{Shan}}$). Consequently, the formulation in Eq. (3) is inefficient, since it requires the calculation of $I_{\mathrm{c}}(q_i; R)$ for a large number N_{exp} of q-points.

The computational cost is reduced by smoothing the raw experimental data, thereby merging neighboring $I_{\mathrm{exp}}(q_i)$ q-points within the same

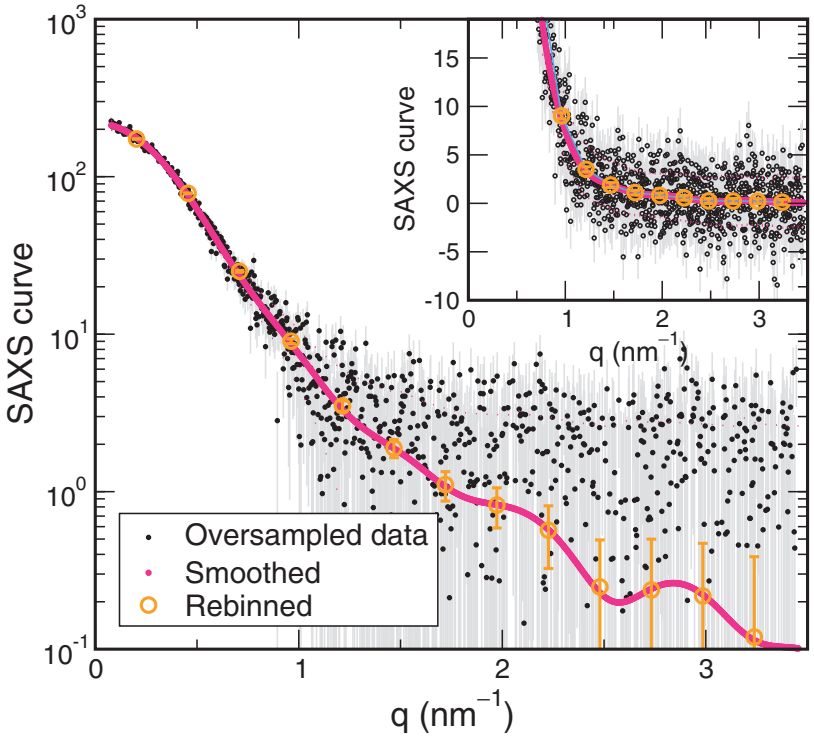

Fig. 2 Smoothing and rebinning the experimental curve to increase the computational efficiency of SAXS-driven MD simulations. *Black dots*: Oversampled experimental SAXS data for bovine serum albumin dimers taken from Jeffries et al. (2016). Curve smoothed with DATGNOM (*pink*) and resampled with DATREGRID, here using 1.5 *q*-points per Shannon bin (Manalastas-Cantos et al., 2021).

Shannon bin into a smoothed curve $\bar{I}_{\mathrm{exp}}(q)$ (Fig. 2). A shell script for smoothing the curve based on the ATSAS module `DATGNOM` (Manalastas-Cantos et al., 2021) is available at https://cbjh.gitlab.io/gromacs-swaxs-docs. Upon smoothing the curve, N_{exp} raw experimental data points are merged into a smooth curve with only N_{Shan} independent features, suggesting that the uncertainties of the smoothed curve follow $\bar{\sigma}^2(q) \approx \sigma^2(q)/n_s$, where $n_s = N_{\mathrm{exp}}/N_{\mathrm{Shan}}$ is the number of experimental points per Shannon bin. Let $\Delta I(q_i) := I_c(q_i; R) - (fI_{\mathrm{exp}}(q_i) + c)$ denote the residuals between experimental and calculated curve and

$$\chi^2 = \sum_{i=1}^{N_{\mathrm{exp}}} \left(\frac{\Delta I(q_i)}{f\sigma(q_i)} \right)^2, \tag{4}$$

such that $E_{\mathrm{exp}} = f_c k_B T \chi^2/2$. Let us further decompose the residuals $\Delta I_i(q_i) = \Delta \bar{I}_i + \delta I_i$ into contributions (i) relative to the smoothed curve $\Delta \bar{I}_i$, which can be fitted by adjusting the biomolecular structure R, and

(ii) owing to statistical noise in the data δI_i, which cannot be fitted. Then, χ^2 can be rewritten as

$$\chi^2 = \sum_{i=1}^{N_{\text{exp}}} \left(\frac{\Delta \bar{I}(q_i)}{f\sigma(q_i)} \right)^2 + \sum_{i=1}^{N_{\text{exp}}} \left(\frac{\delta I(q_i)}{f\sigma(q_i)} \right)^2 + 2 \sum_{i=1}^{N_{\text{exp}}} \frac{\Delta \bar{I}(q_i)\ \delta I(q_i)}{[f\sigma(q_i)]^2} \quad (5)$$

The third term in Eq. (5) vanishes approximately because the noise $\delta I(q_i)$ is symmetrically distributed around zero. The second term adds a constant offset of ≈ 1 per experimental data point, i.e., a total offset of N_{exp} to χ^2 that cannot be fitted. The first term quantifies the deviation with respect to the smoothed curve and contains all the relevant structural information. Using that $\Delta \bar{I}(q_i)$ and $\sigma(q_i)$ are approximately constant within each Shannon bin, the first term of Eq. (5) simplifies to a sum over the Shannon bins,

$$\chi^2 \approx \sum_{b=1}^{N_{\text{Shan}}} \frac{\Delta \bar{I}(q_b)^2}{f^2 \bar{\sigma}(q_b)^2} + N_{\text{exp}}\,, \quad (6)$$

where we used $\bar{\sigma}^2(q) \approx \sigma^2(q)/n_s$. Hence, up to a structurally irrelevant constant offset N_{exp} owing to experimental noise, χ^2 and E_{exp} can be evaluated using only the N_{Shan} intensities and errors of the smoothed curve. In case that more than N_{Shan} q-points are used to evaluate E_{exp}, a correcting prefactor is needed and we arrive at

$$E'_{\text{exp}} \approx \frac{1}{2} f_c\ k_B T \frac{N_{\text{Shan}}}{N_{\text{used}}} \sum_{i=1}^{N_{\text{used}}} \frac{[I_c(q_i; R) - (f\bar{I}_{\text{exp}}(q_i) + c)]^2}{[f\bar{\sigma}(q_i)]^2}, \quad (7)$$

where N_{used} may be chosen in the range of 1–2 N_{Shan}. Evidently, using the formulation in Eq. (7) greatly improves the computational efficiency of SAXS-driven MD simulations as it requires ~30 to 300 times fewer SAXS intensity calculations as compared to the formulation in Eq. (3).

2.3 Accounting for systematic and calculated errors

SAXS data collected with single-photon counting detectors may be subject to tiny statistical errors $\sigma_{\text{exp}}(q)$ at small scattering angles. Consequently, the overall uncertainty of the data may be dominated by unknown systematic errors, for instance owing to imprecise buffer subtraction or minor undetected radiation damage. When applying only the tiny statistical errors, the SAXS-derived energies take spuriously large values at small angles due to the $1/\bar{\sigma}(q_i)^2$ term, which would further propagate into large SAXS-derived forces.

Such problems are avoided by modeling of systematic errors. We previously modeled systematic errors via an uncertainty $\delta\rho_{\mathrm{buf}}$ of 0.1% to 1% of the buffer density (Chen & Hub, 2015). The uncertainty $\delta\rho_{\mathrm{buf}}$ propagates into an uncertainty $\sigma_{\mathrm{buf}}(q)$ of the SAXS curve, which is large at low angles and small at wide angles. Consequently, $\sigma_{\mathrm{buf}}(q)$ dominates the overall uncertainty at low angles and avoids spuriously large SAXS-derived forces, whereas $\bar{\sigma}(q)$ dominates at wide angles. Apart from the systematic errors, GROMACS-SWAXS takes statistical errors of the calculated curve $\sigma_{\mathrm{c}}(q_i)$ into account (Fig. 3). This is implemented by replacing the errors $\bar{\sigma}$ in Eq. (7) with $\sigma_{\mathrm{tot}}^2 = f^2\bar{\sigma}^2 + \sigma_{\mathrm{buf}}^2 + \sigma_{\mathrm{c}}^2$, leading to the final SAXS-derived energy applied by GROMACS-SWAXS:

$$E_{\mathrm{exp}}^{\mathrm{f}} \approx \frac{1}{2} f_c \, k_B T \frac{N_{\mathrm{Shan}}}{N_{\mathrm{used}}} \sum_{i=1}^{N_{\mathrm{used}}} \frac{\left[I_{\mathrm{c}}(q_i;R) - (f\bar{I}_{\mathrm{exp}}(q_i)+c)\right]^2}{f^2\bar{\sigma}^2(q_i)+\sigma_{\mathrm{buf}}^2(q_i)+\sigma_{\mathrm{c}}^2(q_i)}, \tag{8}$$

2.4 On-the-fly averaging of the calculated SAXS curve

Explicit-solvent SAXS curve predictions require averaging over multiple MD frames for computing a SAXS curve. The number of frames required for obtaining a converged SAXS curve depends on the contrast between the biomolecule and the pure buffer; the smaller the electron density contrast,

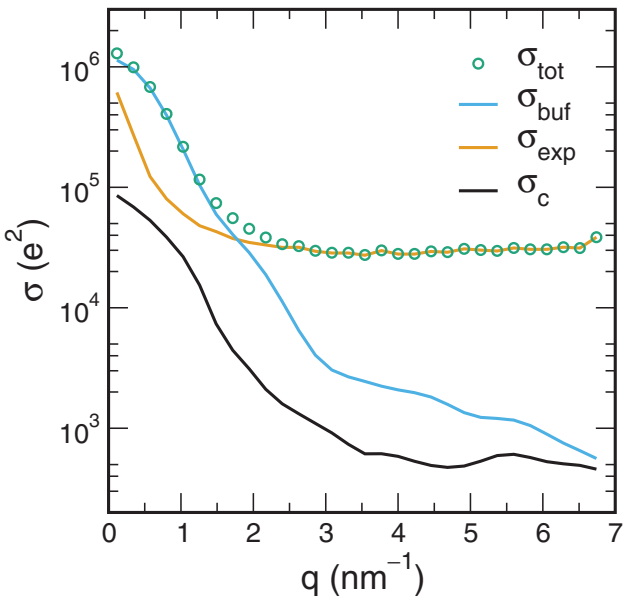

Fig. 3 Typical errors applied during SAXS-driven MD. At small angles, systematic errors σ_{buf} (*blue*) dominate the total uncertainty σ_{tot} (*green circles*), here modeled by a buffer density uncertainty of 1%. At wide angles, experimental errors dominate (*orange*). Calculated statistical errors are typically insignificant (*black*).

the more frames are required to converge the SAXS curve. During SAXS-driven MD simulations discussed here, the SAXS curve is therefore computed as a running temporal average with a memory kernel that decays exponentially into the past. The on-the-fly average of a quantity X at simulation time t is then given by

$$\langle X \rangle(t) = \mathcal{N}^{-1} \int_0^t X(t') \, \mathrm{e}^{-(t-t')/\tau} dt', \tag{9}$$

where τ is the memory time, typically chosen between 50 and 200 ps, and $\mathcal{N}$ is a normalization constant. Likewise, on-the-fly averaging of the scattering amplitudes of the biomolecular and the pure-solvent system using Eq. (9) provide an on-the-fly averaged SAXS curve (Chen & Hub, 2015).

Implementing the coupling to the experimental SAXS curve with an on-the-fly averaged calculated curve has two key advantages: first, thermal fluctuations on the time scale 1–2 τ such as solvent, side chain, or loop fluctuations are taken into account before comparing the calculated with the experimental curve in Eq. (8). Such thermal fluctuations may significantly influence the SAXS curve of proteins at moderate scattering angles ($q > 0.25$ Å^{-1}) (Chen & Hub, 2014; Moore, 2014; Tiede, Zhang, & Seifert, 2002). Second, there is no need to compute SAXS intensities every MD step, but instead one SAXS update every ~0.5 ps is sufficient. The update interval together with the memory time must be chosen such that the SAXS curves converge within τ. The longer update interval renders SAXS-driven MD simulation computationally efficient with an overhead of only 5–20% relative to unbiased simulations. However, owing to the on-the-fly average, the dynamics are not conservative because the energy $E^{\mathrm{f}}_{\mathrm{exp}}$ depends not only on the current but also on previous conformations. Hence, a reasonably tight temperature coupling scheme is required to avoid energy drifts.

2.5 SAXS-derived forces applied during MD

SAXS-derived forces applied in the MD simulations are given by the negative gradients of the SAXS-derived energy,

$$\mathbf{F}_k = -\nabla_k E^{\mathrm{f}}_{\mathrm{exp}}(R, D) \tag{10}$$

$$= -f_c \, k_B T \frac{N_{\mathrm{Shan}}}{N_{\mathrm{used}}} \sum_{i=1}^{N_{\mathrm{used}}} \frac{I_{\mathrm{c}}(q_i; R(t)) - (f \bar{I}_{\mathrm{exp}}(q_i) + c)}{\sigma_{\mathrm{tot}}(q_i)^2} \nabla_k I_{\mathrm{c}}(q_i; R(t)), \tag{11}$$

where $\nabla_k I_c(q_i; R(t))$ denotes the on-the-fly averaged gradients of the SAXS intensities with respect to the coordinates of atom k. For large biomolecular systems, storing these gradients may require several Gigabytes of memory (Chen & Hub, 2015).

GROMACS-SWAXS computes the forces $\mathbf{F}_k$ and the gradients $\nabla_k I_c(q_i;R(t))$ only for the solute atoms but not for the solvent atoms. Therefore, in addition to the derivatives of the I_c with respect to the solute coordinates, a correction factor

$$f_{\text{contrast}} = \frac{\rho_{\text{solu}} - \rho_{\text{solv}}}{\rho_{\text{solu}}} \tag{12}$$

is applied, where ρ_{solu} and ρ_{solv} denote the average solute and solvent densities, respectively. Here, density refers to the scattering length density, that is, to the electron density in SAXS or to the neutron scattering length density in SANS. As a numerical example, the factor f_{contrast} accounts for the fact that, upon moving a protein domain with density $\rho_{\text{solu}} = 440$ e nm^{-3} inside solvent with density $\rho_{\text{solv}} = 334$ e nm^{-3}, only the contrast of 106 e nm^{-3} is moved. Consequently, change of $I_c(q_i)$ upon a domain movement in mobile water is reduced by the factor $f_{\text{contrast}} = 0.24$ as compared to a domain movement at fixed water positions.

When coupling to SANS data measured at a large D_2O concentration, the contrast and the factor f_{contrast} may even become negative because a D_2O buffer exhibits a larger neutron scattering length density than proteins. Consequently, upon moving a protein domain to the left, the contrast may move to the right; such effect is correctly taken into account by a negative factor f_{contrast}.

2.6 Protocol A

To carry out SWAXS-driven MD simulations with the GROMACS-SWAXS implementation published at https://gitlab.com/cbjh/gromacs-swaxs, the following protocol is recommended:

1. Download and compile GROMACS-SWAXS, following the installation instructions of official GROMACS. Alternatively, GROMACS-SWAXS can be easily installed using Spack, which is available at many high-performance computing centers.
2. Setup the MD simulation system following freely available GROMACS tutorials. Make sure to use an approximately 1 nm larger simulation box compared to common simulations via the GROMACS module `gmx editconf -d`.

3. First compute a SAXS curve from an equilibrium simulation using the rerun functionality of the GROMACS-SWAXS `gmx mdrun` module, following tutorials at https://cbjh.gitlab.io/gromacs-swaxs-docs and as described in chapter "Predicting solution scattering patterns with explicit-solvent molecular simulations" by Chatzimagas and Hub in Part A of this monograph. Computing a SAXS curve involves the setup of a pure-solvent simulation systems, definition of the atomic form factors with the `gmx genscatt` module, and building of the envelope with `gmx genenv`, which defines the solvent region contributing to the SAXS curve. From visual inspection of the SAXS curve, is it plausible that a conformational transition of the biomolecule explains the deviation between the calculated and the experimental curve?
4. Generate the run-input (tpr) file with the `gmx grompp` module. For a typical SAXS-driven MD simulation, the MD parameter (mdp) file may look as follows:

```
; read scattering info from topology
define           = -DSCATTER
; turn on SAXS calc. and, optionally, multiple SANS calc.
scatt-coupl      = xray [neutron neutron ...]
; solute group
waxs-solute      = Protein ; or Protein-Masses
; solvent group
waxs-solvent     = Water_and_ions
; rotational fit group as defined with gmx genenv
waxs-rotfit      = C-alpha
; memory time for on-the-fly average
waxs-tau         = 200
; turn on SAXS energy E_exp slowly, e.g. over 5ns
waxs-t-target    = 5000
; define coupling potential type. Alternative: log
waxs-potential   = linear
; use experiental, calculated and systematic errors
waxs-weights     = exp+calc+solvdens
; force constant, use 1-5. fc=1 enables Bayesian interpretation
waxs-fc          = 1
; fit scale f and offset c of experimental curve
waxs-Iexp-fit    = scale-and-offset
; update on-the-fly averaged SAXS curve every 0.5ps with
dt=0.002
waxs-nstcalc     = 250
```

```
; nr of q-values, N_used. Use approx. 1.5 N_Shan
waxs-nq          = 15
; qmin and qmax in nm^(-1)
waxs-startq      = 0
waxs-endq        = 5
; nr of q-vectors for orientational average, use ~0.2*(D*qmax)^2
waxs-nsphere     = 500
; density of solvent (e/nm3), used for a precise buffer subtraction
waxs-solvdens    = 334
; relative uncertainty of the solvent density, used to
; estimate systematic errors. Use 0.1% to 0.5%
waxs-solvdens-uncert = 0.005
; D2O concentration (e.g. 0% and 100%) for each SANS calc.
waxs-deuter-conc = [0 1]
```

5. Smooth the experimental target curve. Upon smoothing the curve, the errors are reduced by a factor of $n_s^{1/2}$, where n_s is the number of experimental data points per Shannon bin. A Shell script for this purpose is available at https://cbjh.gitlab.io/gromacs-swaxs-docs. The shell script requires installation of the ATSAS software (Manalastas-Cantos et al., 2021).
6. Finally, specify the envelope files with environment variables and run the SAXS-driven MD:

```
export GMX_ENVELOPE_FILE=envelope.dat
export GMX_WAXS_FIT_REFFILE=envelope-ref.gro
sw=/path/to/pure-solvent.tpr
fw=/path/to/pure-solvent.xtc
gmx mdrun -s topol.tpr -fw $fw -sw $sw -is target.xvg ...
```

   Here, `pure-solvent.tpr` and `pure-solvent.xtc` are the run-input and trajectory files of the pure-solvent systems used for computing the buffer subtraction. These have already been set up for computing the SAXS curve from an equilibrium simulation in step 3.
7. To analyze the simulation, visualize it in a molecular viewer and validate that the conformation is reasonable. Inspect in `waxs_final.xvg` whether the simulation was capable of finding a conformation that is compatible with the target SAXS curve. Inspect the SAXS-derived energy $E^{\mathrm{f}}_{\mathrm{exp}}$, which is stored in the energy (edr) file and available via the `gmx energy` command. Validate that the energy is in the range of several k_BT after the structure has been refined. The contributions of individual q-points to $E^{\mathrm{f}}_{\mathrm{exp}}$, available in `waxs_pot.xvg`, may reveal q-regions that could not be explained.

8. Since `waxs_final.xvg` represents the final on-the-fly average of the calculated curve, it represents only the final 1–2 τ (~200 ps) of the simulation. Use the rerun functionality of mdrun (`gmx mdrun -rerun traj.xtc`) to compute the SAXS curve that is uniformly averaged over a longer part of the SAXS-driven simulation, for instance over the final 10 ns. A uniform average is enabled with the mdp option `waxs-tau = 0`. Use more q-points, such as `waxs-nq = 100`.

3. SAXS-driven MD as a tool for Bayesian inference of molecular structures

3.1 Posterior, likelihood, and prior distributions

Following pioneering work by Rieping, Habeck, and coworkers on refinement against NMR data (Habeck, Rieping, & Nilges, 2006; Rieping, Habeck, & Nilges, 2005), SAXS-driven MD simulations have been reformulated as tool for Bayesian inference of biomolecular structures from experimental data. Accordingly, the goal of structure refinement is to find the conditional probability $P(R, \theta|D, K)$ that quantifies the plausibility for the biomolecular structure R in the light of the experimental data D and prior physical knowledge K. In the context of SAXS-guided modeling, the data D is given by the experimental curve $I_{\text{exp}}(q)$ and its errors, whereas the prior knowledge K represents the MD force field, the starting conformation of the simulation, and other prior experience. The symbol θ summarizes so-called nuisance parameters such as the unknown fitting parameters or unknown systematic errors, which must be estimated simultaneously with the structure.

In Bayesian inference, $P(R, \theta|D, K)$ should *not* be interpreted as probabilities of random events, as common in the "frequentist interpretation" of probability theory, but instead as a quantification of our state of knowledge and ignorance. A wide distribution $P(R, \theta|D, K)$ implies that many different conformations R are compatible with D and K, implying a high degree of ignorance and large confidence intervals on R. By the same token, a narrow $P(R, \theta|D, K)$ implies that only few structures are plausible in the light of D and K, implying precise knowledge of R and small confidence intervals on R. Hence, computing $P(R, \theta|D, K)$ provides rigorous confidence intervals for the refined structure founded in probability theory.

The conditional probability is evaluated using Bayes' theorem,

$$P(R, \theta|D, K) \propto L(D|R, \theta, K)\, \pi(R|K)\, \pi(\theta|K), \tag{13}$$

where $\pi(R|K)$ and $\pi(\theta|K)$ denote the prior distributions of conformations and of the nuisance parameters, respectively. $L(D|R, \theta, K)$ is the likelihood that the data D was measured given that the structure R was present in the experiment and given a specific set of nuisance parameters θ. The distribution $P(R, \theta|D, K)$ is called *posterior distribution*. When running MD simulations, the prior $\pi(R|K)$ is naturally given by the Boltzmann distribution

$$\pi(R|K) \propto \mathrm{e}^{-V_{\mathrm{FF}}(R)/k_B T} \tag{14}$$

with the MD force field $V_{\mathrm{FF}}(R)$, i.e., by the ensemble obtained before incorporating the data D. A reasonable choice for $\pi(\theta|K)$ is less obvious; if no prior information on the nuisance parameters is available, a noninformative prior such as a flat or a Jeffreys prior may be used. It is advisable to test multiple priors in order to exclude that the conclusions depend on the choice of the priors. For instance, testing multiple choices of $\pi(R|K)$ corresponds to running simulations with difference force fields V_{FF}.

Assuming Gaussian independent errors σ_{tot}, the likelihood function is

$$\begin{aligned} L(D|R,\theta,K) &\propto \prod_{i=1}^{N_{\mathrm{exp}}} \exp\left(-\frac{[I_{\mathrm{c}}(q_i;R) - I'_{\mathrm{exp}}(q_i;\theta)]^2}{2[f\sigma(q_i;\theta)]^2}\right) \\ &= \exp\left(-\frac{1}{2}\sum_{i=1}^{N_{\mathrm{exp}}} \frac{[I_{\mathrm{c}}(q_i;R,\theta) - I'_{\mathrm{exp}}(q_i;\theta)]^2}{[f\sigma(q_i;\theta)]^2}\right), \end{aligned} \tag{15}$$

Here, the $I'_{\mathrm{exp}}(q_i;\theta) = f\, I_{\mathrm{exp}}(q_i) + c$ is the experimental data adjusted by the scale f and the offset c, which appear as nuisance parameters (cf. Eq. 3). In GROMACS-SWAXS, the likelihood function is modified twofold relative to Eq. (15). First, as described above, the raw experimental intensities (I_{exp}, σ) and errors are replaced with the smoothed curve ($\bar{I}_{\mathrm{exp}}$, $\bar{\sigma}$), thereby replacing the sum over N_{exp} values with a sum over N_{Shan} values. Second, the errors are augmented with the calculated and systematic errors, leading to the final likelihood function:

$$L^{\mathrm{f}}(D|R,\theta,K) \propto \exp\left(-\frac{1}{2}\frac{N_{\mathrm{Shan}}}{N_{\mathrm{used}}}\sum_{j=1}^{N_{\mathrm{used}}} \frac{(I_{\mathrm{c}}(q_j;R) - \bar{I}'_{\mathrm{exp}}(q_j;\theta))^2}{f^2\bar{\sigma}^2(q_i) + \sigma^2_{\mathrm{buf}}(q_i;\theta) + \sigma^2_{\mathrm{c}}(q_i)}\right) \tag{16}$$

Hence, L^{f} contains three nuisance parameters $\theta = \{f, c, \delta\rho_{\mathrm{buf}}\}$.

The posterior distribution $P(R, \theta|D, K)$ cannot be computed analytically but is instead obtained by importance sampling. Here, Newtonian dynamics as implemented by MD simulations are used for sampling the conformations.

By taking the negative logarithm of the posterior, we turn the probability functions into energy terms,

$$\begin{aligned} E &= -k_B T \ln P(R,\theta|D,K) \\ &= V_{\mathrm{FF}}(R) + E^{\mathrm{f}}_{\mathrm{exp}}(R, D, \theta)|_{f_c=1} - k_B T \ln \pi(\theta|K), \end{aligned} \tag{17}$$

where we used Eqs. (8), (13), (14), and (16). Several aspects of this results are notable:

- The original SAXS-derived energy Eq. (8) required the choice of a force constant f_c that weights the experimental data relative to the force field V_{FF}. Using Bayesian inference, in contrast, no force constant is needed because the weight of the experimental data is fully determined by probability theory. In practice, it may be useful to first drive a conformational transition with a larger f_c and, in a follow-up simulation, sample the posterior with $f_c = 1$.
- Sampling the posterior $P(R, \theta|D, K)$ has been implemented using Gibbs sampling. Accordingly, nuisance parameters θ are sampled with Metropolis Monte-Carlo at fixed conformations R, followed by Newtonian dynamics of R at fixed θ, and so on.
- Sampling the fitting parameters f and c is not required because they can be marginalized out analytically at the level of the likelihood when assuming flat priors for f and c, i.e., $\pi(f|K) = \pi(c|K) = \text{const}$. Then, the likelihood in Eq. (16) is replaced with $\widetilde{L}(D|R,\delta\rho_{\mathrm{buf}},K) = \int L(D|R,\theta,K)\, \mathrm{d}f\, \mathrm{d}c$, thereby taking "all possible" values of f and c into account. Evaluating this integral shows that $\widetilde{L}$ takes the same form as L, except that f and c are replaced with their maximum-likelihood estimates f_{ml} and c_{ml} (Shevchuk & Hub, 2017).

3.2 Bayesian treatment of systematic errors at small angles

GROMACS-SWAXS models systematic errors at small angles via an uncertainty of the buffer density $\delta\rho_{\mathrm{buf}}$ (see above and Fig. 3). In Bayesian SAXS-driven MD, $\delta\rho_{\mathrm{buf}}$ can be treated as one of the nuisance parameters θ (Shevchuk & Hub, 2017). Accordingly, the relative uncertainty $\delta\rho_{\mathrm{buf}}$ is sampled simultaneously with the structure R to obtain a joint posterior distribution over structures and buffer density uncertainties $P(R, \delta\rho_{\mathrm{buf}}|D, K)$. GROMACS-SWAXS applies a Gaussian prior distribution for $\delta\rho_{\mathrm{buf}}$. Let $r_{\mathrm{buf}} := \delta\rho_{\mathrm{buf}}/\rho_{\mathrm{buf}}$, then the prior is

$$\pi(r_{\mathrm{buf}}|K) \propto \exp\left(-r^2_{\mathrm{buf}}/2\epsilon^2_{\mathrm{buf}}\right) \tag{18}$$

where ρ_{buf} is the solvent density and ϵ_{buf} is given with the mdp parameter `waxs-solvdens-uncert`, typically set between 0.1% and 1%. This algorithm detects automatically whether the experimental data is biased by systematic errors at small angles. Namely, if the MD force field permits a conformational transition that explains the experimental data $I_{\mathrm{exp}}(q)$, the posterior of $\delta\rho_{\mathrm{buf}}$ would peak at small values, indicating that systematic errors are not required to explain the experimental data (Fig. 4, blue solid). In contrast, if the MD simulations does not find a conformation that explains I_{exp} at small angles, the posterior of $\delta\rho_{\mathrm{buf}}$ would peak at larger values, indicating that significant systematic errors are plausible in the light of the data and the force field (Fig. 4, orange dashed).

Clearly, treatment of systematic errors in SAXS-based modeling is still underdeveloped. To harvest increasingly finer details in SAXS data, it will

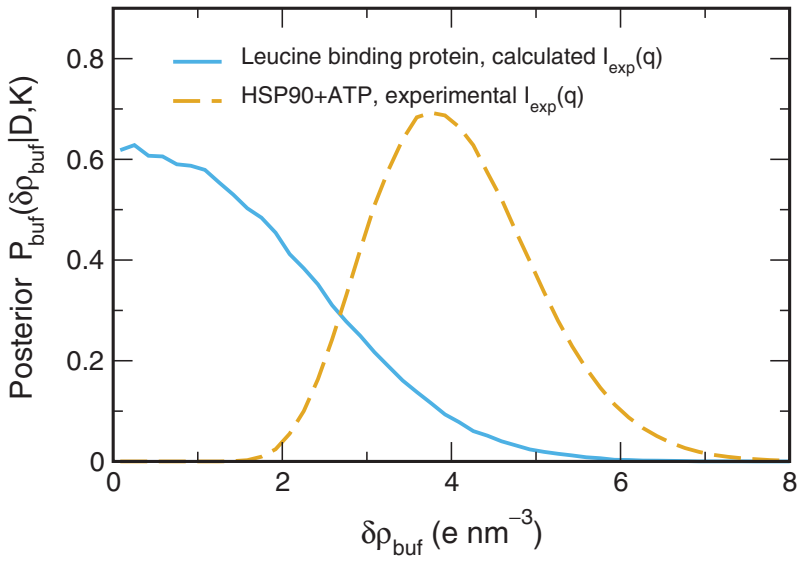

Fig. 4 Example of posterior distributions over the uncertainty of the buffer density $\delta\rho_{\mathrm{buf}}$ taken from SAXS-driven MD simulations of leucine binding protein (LBP, *solid curve*) or HSP90 bound to ATP (*dashed curve*). For LBP, the posterior peaks at $\delta\rho_{\mathrm{buf}} = 0$; hence, systematic errors are not required to explain the data, but they cannot be excluded either. This finding is expected since the SAXS-driven MD simulation were carried out against a calculated target curve without any uncertainties. In contrast, the posterior for HSP90 peaks at larger $\delta\rho_{\mathrm{buf}} = 4$ e nm^{-3}, suggesting that significant systematic errors are strictly required to explain the data. This finding reflects that the SAXS-driven MD was carried out against experimental data with substantial systematic errors at small angles. *Data taken from Shevchuk, R., & Hub, J. S. (2017). Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. PLOS Computational Biology, 13, e1005800.*

be useful to explicitly model other sources of errors, such as a small fraction of aggregated samples. We believe that Bayesian inference will provide a rigorous framework to learn systematic errors simultaneously with the biomolecular structures.

3.3 Protocol B

The Bayesian interpretation of SAXS-driven MD simulations are enabled by using a force constant of unity, and optionally, by sampling $\delta\rho_{\mathrm{buf}}$. The mdp file for a Bayesian SAXS-driven MD simulation should contain, apart from the options described above:

```
waxs-fc= 1
waxs-solvdens-uncert= 0.005
waxs-solvdens-uncert-bayesian= yes
```

The simulation frames in the trajectory of the SAXS-driven MD may be interpreted as samples from a high-dimensional posterior distribution over conformations $P(R|D, K)$. Obtaining posteriors over intuitive properties, such as the center-of-mass distance d_{com} between two domains, is straightforward: d_{com} values may be extracted from the trajectory frames and plotted as a histogram. The histograms are the posterior $P_{\mathrm{com}}(d_{\mathrm{com}}|D, K)$ over the center-of-mass distances, suggesting that the peak position and the width of P_{com} provide the most plausible d_{com} and its uncertainty in the light of the data and the force field. Mathematically, obtaining $P_{\mathrm{com}}(d_{\mathrm{com}}|D, K)$ from $P(R|D, K)$ would involve a marginalization, that is the integration over all other degrees of freedom except d_{com}; however, since the trajectory of the Bayesian SAXS-driven MD contains samples of d_{com}, there is no need carry out a marginalization in practice.

Samples from the posterior over the relative uncertainty of the buffer density, $\delta\rho_{\mathrm{buf}}/\rho_{\mathrm{buf}}$, are written into separate output file `waxs_solvDensUncert.xvg` allowing the calculation of a histogram and, hence, the posterior over the uncertainty of the buffer density.

4. Maximum-entropy ensemble refinement against SAXS data

4.1 Theoretical background

Since SAXS is a solution method, experimental SAXS intensities represent the average over a structural ensemble. For structurally stable proteins, the ensemble may be approximated by a single, most prominent conformation,

enabling the use of structure refinement methods described above. In contrast, for solutes that adopt a heterogeneous ensemble, the ensemble is adequately represented by a large number of conformations or by conformational distributions. Typical examples are intrinsically disordered proteins (IDPs) and proteins with disordered regions, domains connected with flexible linkers, or dynamic soft-matter complexes. Upon refining such heterogeneous ensembles against experimental SAXS data $I_{\mathrm{exp}}(q)$, the SAXS curves computed from the individual conformations may differ from $I_{\mathrm{exp}}(q)$, and only the ensemble-averaged computed curve should agree with the data.

However, ensemble refinement against SAXS data is an ill-posed problem because many different ensembles would agree with $I_{\mathrm{exp}}(q)$, even if the conformational space is restrained with an all-atom force field. Two strategies have been put forward for choosing a justified refined ensemble from all the ensembles that satisfies the data (Ravera, Sgheri, Parigi, & Luchinat, 2016). Following the strategy of maximum parsimony, the aim is to explain the data with as few conformations as possible. Such approach is most justified if the biomolecule adopts a few well-defined conformational states. The second approach is founded in statistical physics and is based on Jaynes' maximum-entropy principle (Boomsma, Ferkinghoff-Borg, & Lindorff-Larsen, 2014; Cesari, Reißer, & Bussi, 2018; Hermann & Hub, 2019; Jaynes, 1957). According to Jaynes, we should choose the ensemble distribution with the greatest uncertainty (or with the least information) that satisfies a given set of constraints. In the context of structure refinement, the constraints are our requested agreement with the experimental data. The principle is satisfied by finding an ensemble distribution that maximizes the Shannon entropy

$$S(p) = -\sum_i p(R_i) \ln p(R_i) \tag{19}$$

under the given constrains.

In ensemble refinement, however, we are typically interested in refining a prior ensemble from a free MD simulation against experimental data, suggesting that it is useful to maximize the *relative entropy* between the unbiased and the refined ensemble (Caticha, 2004). Because the relative entropy is the negative of the Kullback–Leibler divergence $D_{\mathrm{KL}}(p_1|p_0)$ (Kullback & Leibler, 1951), maximizing the relative entropy implies that we should find a refined ensemble distribution $p_1(R_i)$ that minimizes

$$D_{\mathrm{KL}}(p_1|p_0) = \sum_i p_1(R_i) \ln \frac{p_1(R_i)}{p_0(R_i)} \tag{20}$$

under the given constrains, where $p_0(R_i)$ is the prior ensemble distribution. Taking the prior from an unbiased MD ensemble, the aim is to find an updated ensemble that is (i) in agreement with the data and (ii) updated as minimally as possible with respect to the unbiased prior ensemble. In other words, the ensemble is only updated as strictly needed to explain the data, while any bias that is not supported by the data is avoided. Specifically, the formulation assures that the ensemble is not updated if the prior ensemble is already in agreement with the data.

The minimization problem can be solved with the help of Lagrangian multipliers, where one multiplier is required for each experimental constraint (Pitera & Chodera, 2012). However, the Lagrangian multipliers must be optimized in an iterative manner, which may be tedious in presence of a larger number of experimental constraints (Boomsma et al., 2014). An alternative for implementing a minimal bias is the parallel-replica approach. Here, several copies of the system (replicas) are simulated in parallel, and only the calculated data averaged over the replicas is restrained to the data with a harmonic restraint (Fig. 5). Roux and Weare (2013) and Cavalli, Camilloni, and Vendruscolo (2013) showed that the replica-averaged harmonic restraint imposes a minimal bias in the limit of a large number of replicas.

4.2 Parallel-replica ensemble refinement against SAXS data

Parallel-replica refinement against SAXS data is illustrated for an IDP in Fig. 5. First, the SAXS intensity is averaged over the replicas intensities $I_c(q_i, R_\alpha)$,

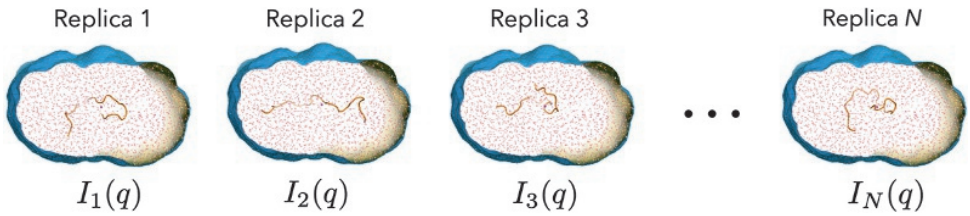

Fig. 5 Illustration of parallel-replica ensemble refinement. *N* replicas are simulated simultaneously, each providing a calculated curve $I_1,\ldots,I_N$. Coupling the replica-averaged SAXS curve to the experiment with a harmonic restraint leads to the maximum entropy ensemble.

$$\bar{I}_c(q_i; R_1,\ldots,R_{N_{\text{rep}}}) = \frac{1}{N_{\text{rep}}} \sum_{\alpha=1}^{N_{\text{rep}}} I_c(q_i, R_\alpha), \tag{21}$$

where α is the replica index and N_{rep} is the number of replicas. Then, the systems are coupled to the data with a harmonic restraint similar to Eq. (8), except that the SAXS curve from a single simulation is substituted by the replica-averaged curve:

$$E_{\text{exp}}\left(R_1,\cdots,R_{N_{\text{rep}}};I_{\text{exp}}\right) = \frac{1}{2} N_{\text{rep}} f_c k_B T \frac{N_{\text{Shan}}}{N_{\text{used}}} \sum_j^{N_{\text{used}}} \frac{\left[\bar{I}_c\left(q_j;R_1,\cdots R_N\right) - \bar{I}'_{\text{exp}}\left(q_j\right)\right]^2}{f^2\,\bar{\sigma}^2\left(q_j\right)} \tag{22}$$

Here, $\bar{I}'_{\text{exp}}$ denotes the smoothed experimental curve adjusted by the fitting parameters f and c, as used above. The biasing energy is multiplied with N_{rep}, such that the factor cancels with $1/N_{\text{rep}}$ in Eq. (21) when taking the derivatives with respect to atomic coordinates, as done for computing the SAXS-derived forces (Hummer & Köfinger, 2015).

4.3 Choosing the number of replicas

The number of replicas that are required to follow the maximum entropy principle depends on the system (Boomsma et al., 2014; Hermann & Hub, 2019). A possible strategy for finding a good value for N_{rep} is to investigate distributions $h(\xi)$ of a few important degrees of freedom ξ, such as distribution of the radius of gyration of an IDP or of the moments of inertia of a soft-matter complex. Accordingly, the Kullback–Leibler divergence $D_{\text{KL}}(h_1|h_0)$ between the biased distribution $h_1(\xi)$ and unbiased distribution $h_0(\xi)$ may be plotted vs the number of replicas. A sufficient value of N_{rep} would be indicated by a plateau region of such plot.

A disadvantage of D_{KL} is its numerical instability; namely, since p_0 appears in the denominator of Eq. (20), D_{KL} is unstable if some conformations of the biased ensemble p_1 were hardly sampled in the unbiased distribution p_0. A numerically more stable alternative is given by the Jensen–Shannon divergence, which may be considered as a smoothed and symmetrized version of D_{KL},

$$D_{\text{JS}}(h_1|h_0) = [D_{\text{KL}}(h_0|h_M) + D_{\text{KL}}(h_1|h_M)]/2, \tag{23}$$

where $h_M = (h_0 + h_1)/2$ is the average of h_0 and h_1 (Hermann & Hub, 2019). Another useful measure is given by the entropy of biased distributions h_1, $S[h_1] = \int h_1(\xi) \ln h_1(\xi) \mathrm{d}\xi$. If the number of replicas is too low, the ensemble is still overly biased, which may lead to overly narrow ensemble and overly narrow distributions $h_1(\xi)$. Hence, it is useful to plot the entropy (or even simply the width) of $h_1(\xi)$ versus the number of replicas. For the refinement of an IDP ensemble we found previously that using only 4–8 replicas were sufficient. This finding is likely explained by the fact that, when using explicit-solvent SAXS predictions, $I_c(q_i, R_\alpha)$ represent on-the-fly averages with a memory time of 50–200 ps. Hence, even the SAXS curves from individual replicas represent some conformation heterogeneity, explaining why only few replicas are sufficient to represent the heterogeneous overall shape of the IDP ensemble that is encoded by the data.

4.4 Protocol C

Multireplica SAXS-driven MD simulations are set up similar to the single-replica refinement described above.

1. Compile GROMACS-SWAXS with MPI support using
   ```
   cmake -DGMX_MPI=ON ...
   ```
2. When using four replicas, for instance, use the following mdp options:
   ```
   waxs-ensemble-type    = maxent-ensemble
   waxs-ensemble-nstates = 4
   waxs-scale-i0         = no ; yes with small contrast, e.g. with IDP
   gen-vel               = yes
   ```
   With the `gen-vel = yes` option, the replicas start with different initial velocities, providing independent simulations. For solutes with a small contrast, such as an IDP, the forward scattering intensity $I_c(q = 0)$ may not converge within the memory time τ. To greatly accelerate the convergence, $I_c(q = 0)$ may be fixed to the forward intensity of the target curve by adding a small constant density to the solvent, turned on with the mdp option `waxs-scale-i0 = yes`. Prepare one tpr file for each replica, and place the tpr files into different subdirectories 000, 001, etc:
   ```
   gmx grompp -f maxent.mdp -o 000/topol.tpr
   gmx grompp -f maxent.mdp -o 001/topol.tpr etc.
   ```
3. Run the multireplica simulation with the `-multidir` functionality of mdrun:

```
sw=/path/to/pure-solvent.tpr
fw=/path/to/pure-solvent.xtc
target=$(realpath Itarget_trans.xvg)
export GMX_ENVELOPE_FILE=$(realpath envelope.dat)
export GMX_WAXS_FIT_REFFILE=$(realpath envelope-ref.gro)
mpiexec -np 4 gmx_mpi mdrun \
-s topol.tpr -sw $sw -fw $fw -is $target -multidir 000 001
```

4. Carry out the same analysis as described above for regular SAXS-driven MD. Now, the file `waxs_final.xvg` contains the final on-the-fly average of the replica-averaged SAXS curve. Carry out a rerun with the trajectories of the four replicas to compute a uniformly averaged SAXS curve for each replica, and henceforth average the calculated curve. Inspect whether the replica-averaged curve agrees with the experimental target curve.
5. Compute distributions $h_1(\xi)$ of interesting observables ξ, combined from all replica trajectories, such as the distribution of the radius of gyration or the secondary structure content of an IDP. Such distributions quantify the heterogeneity of the ensemble.
6. With a new simulation system, redo the simulation with increasing numbers of replicas N_{rep}. Recompute distributions over observables using the same aggregated simulation time (e.g., 1×400 ns, 2×200 ns, 4×100 ns, etc.). Inspect convergence of the entropy (or the width) of the distributions $h_1(\xi)$ as function of N_{rep}.

4.5 Example: Ensemble refinement of a detergent micelle

Fig. 6 presents a multireplica-average ensemble refinement of a DDM detergent micelle using an increasing number of 1–10 replicas. The distributions of moments of inertia as computed from the aggregated simulations strongly depend on the number of replicas. Namely, the distributions from single-copy refinement are overly narrow, indicative of an overly restrained ensemble in violation of the maximum entropy principle (Fig. 6C, black). Using a larger number of replicas, the distributions are wider, reflecting a larger degree of heterogeneity. Critically, all simulations reveal quantitative agreement with the data, even when using only a single replica (Fig. 6B). This demonstrates that agreement with the data does by far not guarantee that the ensemble is correct.

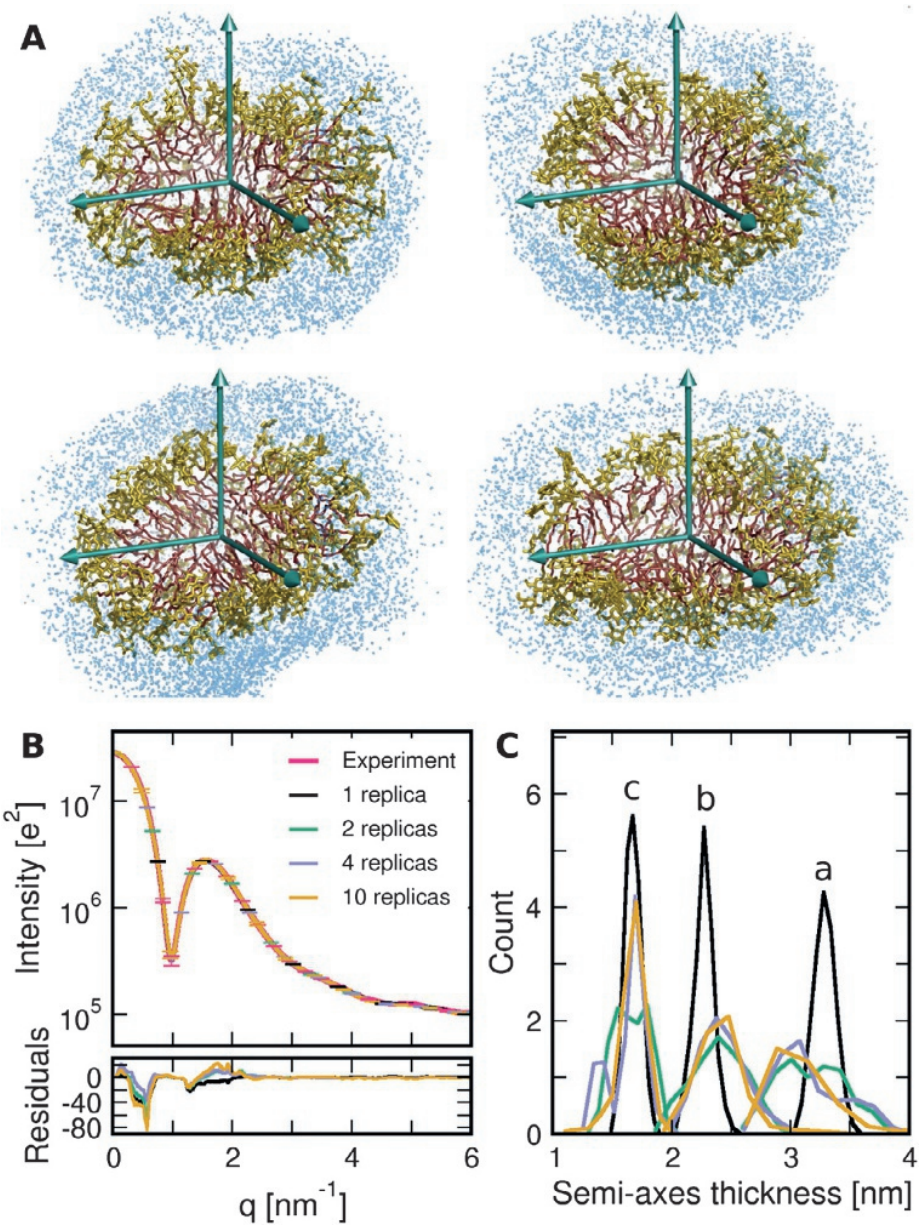

Fig. 6 (A) Parallel-replica ensemble refinement of a DDM detergent micelle. (B) Excellent agreement is found with the data, irrespective of the number of replicas between 1 and 10 ($\chi^2 = 129$, $\chi^2 = 80$, $\chi^2 = 37$, and $\chi^2 = 161$ for 1, 2, 4, and 10 replicas, respectively). (C) However, different numbers of replicas lead to different conformation ensembles, here quantified by the distributions of the micelle core semiaxes a, b, and c, demonstrating that agreement with the data does by far not guarantee that the ensemble is correct. *Reprinted with permission from Ivanović, M. T., Hermann, M. R., Wójcik, M., Pérez, J., & Hub, J. S. (2020). Small-angle X-ray scattering curves of detergent micelles: Effects of asymmetry, shape fluctuations, disorder, and atomic details. The Journal of Physical Chemistry Letters, 11(3), 945–951. https://doi.org/10.1021/acs.jpclett.9b03154. Copyright 2020 American Chemical Society.*

5. Discussion: Conceptual considerations and recommendations

5.1 SAXS-driven MD simulations (should) feel only a weak bias by the SAXS data

Structural data is typically insufficient for defining all degrees of freedom of a biomolecule. This is true not only for SAXS data, but also for data from X-ray crystallography, NMR spectroscopy, or cryoelectron microscopy. To avoid overfitting during structure refinement, structural data is complemented with additional physicochemical information that restraints the biomolecule into realistic conformations. The required amount of additional information critically depends on the information content of the data; the lower information content of the data, the more predictive additional information is needed. For instance, excluding atomic overlaps and restraining chemical bond geometries is often sufficient for the refinement of atomic models against crystallographic data but would be by far insufficient for refinement against SAXS data.

During SAXS-driven MD, the structure is largely imposed by the all-atom force field, which restraints not only chemical geometries but also maintains a proper hydrogen bond network and favorable electrostatic interactions. Overfitting is avoided by using only a small force constant f_c for the SAXS-derived restraints in the order of unity, such that biomolecular dynamics are largely controlled by the force field $V_{\mathrm{FF}}(R)$ whereas the energy E_{exp} only mildly pushes the biomolecule into agreement with the data (see Eqs. 8 and 17). Indeed, $E^{\mathrm{f}}_{\mathrm{exp}}$ takes values of few up to tens of kilojoules per mole, whereas the potential energy contributions from Lennard-Jones or Coulomb interactions are typically in the range of hundred thousands to millions of kilojoules per mole.

Inspection of the SAXS-derived potentials is also advised to prevent overinterpretation of the structural ensemble. The potential should converge over simulation time to values in the range of several k_BT (thermal energies). Caution is advised if individual q-points give rise to high potentials, indicating that parts of the SAXS curve cannot be explained by the simulation ensemble. Generally, it is highly desirable to cross-validate the derived ensemble against independent structural or biochemical data that has not been used during the refinement.

5.2 Accelerating transitions with SAXS data and sampling limitations

SAXS-driven MD simulations have been used to accelerate large-scale conformational transitions of biomolecules, which would require prohibitively long simulation times during unbiased simulations. For instance, we showed that SAXS data may be used to drive large-scale opening transition of the large proteins Hsp90 or ATCase in MD simulations (Chen & Hub, 2015; Shevchuk & Hub, 2017). However, SAXS-driven MD works as an enhanced sampling technique only if the SAXS-derived forces point into the direction of the sought-after conformation transition. This is typically true in the case of large-scale domain motions, in particular if these motions modulate the radius of gyration. In contrast, achieving a complex, nonlinear rearrangements in the simulation such as the folding of an unstructured tail into an α-helix is far more challenging. In such cases, the correct final state might be detectable via a low χ^2 in Eq. (5), but it is unlikely that SAXS-derived forces would accelerate the folding transition. Hence, for guiding complex transitions with SAXS data, SAXS-driven MD may be combined with established enhanced sampling techniques such as simulated tempering or Hamiltonian replica exchange.

5.3 Further analysis

The trajectories obtained from SAXS-driven MD are typically post-processed to provide insight that is not directly accessible from the SAXS data. This includes, for instance, analysis of (i) the heterogeneity of the refined ensembles, (ii) interactions at the atomic level such as secondary structure content or hydrogen bond patterns, (iii) transition pathways, or (iv) structural parameters such as the persistence lengths or end-to-end distances of IDPs. In addition, refined ensembles may provide insight into thermodynamic driving forces by means of free energy calculations, which have so far been underexposed in the context of SAXS-driven MD simulations.

6. Applications

Although SAXS-driven MD simulations have emerged recently, they have provided insight into a range of (bio)molecular systems. To mention a few, parallel-replica ensemble refinement of detergent micelles was used to identify capabilities and limitations of simple geometric models in explaining

SAXS data (Ivanović et al., 2020). WAXS-driven MD solved the salt-dependent solution structures of RNA triplex motifs (Chen, He, Kirmizialtin, & Pollack, 2022), the dynamics of RNA upon ligand or ion binding (He, Henning-Knechtel, & Kirmizialtin, 2022), or a large-scale opening transition of a phytochrome (Björling et al., 2015). In a related approach, combined with metainference and metadynamics, simulations restrained to SAXS data characterized a flexible polyubiquitin chain (Jussupow et al., 2020). Disordered intermediate states of unfolding bovine α-lactalbumin were obtained by restraining MD simulations to time-resolved SAXS data (Hsu, Leshchev, Kosheleva, Kohlstedt, & Chen, 2021). Ongoing efforts in our group involve heterogeneous ensembles of RNA–protein complexes involved in translational control. We expect that more applications will emerge as the implementations are now freely available and widely documented.

7. Summary

SAXS is an increasingly valuable tool of integrative structural biology thanks to the accuracy of data collected at modern SEC-SAXS beamlines and thanks to its structural information content, which is not accessible by other techniques. However, the interpretation of the data is challenged by the low information content of the experimental signals, leading to the risk of overinterpreting the data. This risk is mitigated by using MD simulations that add physicochemical information to the data.

In this chapter, we presented three approaches for refining structural ensembles against SAXS data by integrating the experimental data on-the-fly into all-atom MD simulations:

(i) During SAXS-driven MD, an MD simulation is coupled to experimental SAXS data using harmonic restrains on the data, thereby refining an ensemble that may be approximated by a single, most prominent conformation. The SAXS-driven simulations promote conformational transitions compatible with the data. They are capable of overcoming force field imperfections and sampling limitations of unbiased simulations, given that the transitions are geometrically simple.

(ii) When using SAXS-driven MD as a tool for Bayesian inference of biomolecular structures from a given SAXS curve, the MD simulation samples the sought-after posterior distribution. The posterior quantifies the plausibility of a biomolecular structure in the light of the experimental data and prior physicochemical information, as encoded in the MD force fields. This Bayesian framework may estimate

systematic errors at small angles simultaneously with the structure, which enables one to assess whether systematic errors are required to explain the experimental data.

(iii) The parallel-replica approach allows one to refine heterogeneous ensembles with commitment to the maximum-entropy principle. Here, the SAXS curve averaged over several replicas is coupled to the experimental data using harmonic restrains, such that the updated ensemble is compatible with the data but biased as minimally as possible with respect to the unbiased ensemble.

All approaches require a forward model for calculating the SAXS intensities from the atomistic structures. We discussed structure and ensemble refinement based on explicit-solvent SAXS curve calculations, as used in the WAXSiS method, thereby accurately representing the hydration layer and the excluded solvent (Chen & Hub, 2014; Knight & Hub, 2015). An implementation of the methods described here employing the explicit-solvent SAXS curve calculations is freely available in GROMACS-SWAXS (https://gitlab.com/cbjh/gromacs-swaxs).

Acknowledgments

We thank Milos Ivanovic for preparing Fig. 6A. This study was supported by the Deutsche Forschungsgemeinschaft (HU 1971/3-1).

References

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1*, 19–25.

Bernetti, M., & Bussi, G. (2021). Comparing state-of-the-art approaches to back-calculate SAXS spectra from atomistic molecular dynamics simulations. *The European Physical Journal B*, *94*(9), 180. https://doi.org/10.1140/epjb/s10051-021-00186-9.

Berthaud, A., Manzi, J., Pérez, J., & Mangenot, S. (2012). Modeling detergent organization around aquaporin-0 using small-angle X-ray scattering. *Journal of the American Chemical Society*, *134*(24), 10080–10088.

Björling, A., Niebling, S., Marcellini, M., van der Spoel, D., & Westenhoff, S. (2015). Deciphering solution scattering data with experimentally guided molecular dynamics simulations. *Journal of Chemical Theory and Computation*, *11*, 780–787.

Boomsma, W., Ferkinghoff-Borg, J., & Lindorff-Larsen, K. (2014). Combining experiments and simulations using the maximum entropy principle. *PLOS Computational Biology*, *10*(2), e1003406.

Brosey, C. A., & Tainer, J. A. (2019). Evolving SAXS versatility: Solution X-ray scattering for macromolecular architecture, functional landscapes, and integrative structural biology. *Current Opinion in Structural Biology*, *58*, 197–213. https://doi.org/10.1016/j.sbi.2019.04.004.

Caticha, A. (2004). Relative entropy and inductive inference. In *AIP conference proceedings: Vol. 707* (pp. 75–96). American Institute of Physics.

Cavalli, A., Camilloni, C., & Vendruscolo, M. (2013). Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *The Journal of Chemical Physics*, *138*, 03B603.

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

Chen, P.-c., & Hub, J. S. (2014). Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. *Biophysical Journal*, *107*, 435–447.

Chen, P.-c., & Hub, J. S. (2015). Interpretation of solution X-ray scattering by explicit-solvent molecular dynamics. *Biophysical Journal*, *108*, 2573–2584.

Chen, P.-c., Shevchuk, R., Strnad, F. M., Lorenz, C., Karge, L., Gilles, R., et al. (2019). Combined small-angle X-ray and neutron scattering restraints in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, *15*(8), 4687–4698. https://doi.org/10.1021/acs.jctc.9b00292.

Chen, Y.-L., He, W., Kirmizialtin, S., & Pollack, L. (2022). Insights into the structural stability of major groove RNA triplexes by WAXS-guided MD simulations. *Cell Reports Physical Science*, *3*(7), 100971.

Gorba, C., Miyashita, O., & Tama, F. (2008). Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data. *Biophysical Journal*, *94*(5), 1589–1599.

Gräwert, T. W., & Svergun, D. I. (2020). Structural modeling using solution small-angle X-ray scattering (SAXS). *Journal of Molecular Biology*, *432*(9), 3078–3092. https://doi.org/10.1016/j.jmb.2020.01.030.

Habeck, M., Rieping, W., & Nilges, M. (2006). Weighting of experimental evidence in macromolecular structure determination. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(6), 1756–1761.

He, W., Henning-Knechtel, A., & Kirmizialtin, S. (2022). Visualizing RNA structures by SAXS-driven MD simulations. *Frontiers in Bioinformatics*, *2*, 781949. https://doi.org/10.3389/fbinf.

Hermann, M. R., & Hub, J. S. (2019). SAXS-restrained ensemble simulations of intrinsically disordered proteins with commitment to the principle of maximum entropy. *Journal of Chemical Theory and Computation*, *15*(9), 5103–5115. https://doi.org/10.1021/acs.jctc.9b00338.

Hsu, D. J., Leshchev, D., Kosheleva, I., Kohlstedt, K. L., & Chen, L. X. (2020). Integrating solvation shell structure in experimentally driven molecular dynamics using X-ray solution scattering data. *The Journal of Chemical Physics*, *152*(20), 204115. https://doi.org/10.1063/5.0007158.

Hsu, D. J., Leshchev, D., Kosheleva, I., Kohlstedt, K. L., & Chen, L. X. (2021). Unfolding bovine α-lactalbumin with T-jump: Characterizing disordered intermediates via time-resolved x-ray solution scattering and molecular dynamics simulations. *The Journal of Chemical Physics*, *154*(10), 105101.

Hub, J. S. (2018). Interpreting solution X-ray scattering data using molecular simulations. *Current Opinion in Structural Biology*, *49*, 18–26.

Hummer, G., & Köfinger, J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. *The Journal of Chemical Physics*, *143*(24), 12B634_1.

Ivanović, M. T., Hermann, M. R., Wójcik, M., Pérez, J., & Hub, J. S. (2020). Small-angle X-ray scattering curves of detergent micelles: Effects of asymmetry, shape fluctuations, disorder, and atomic details. *The Journal of Physical Chemistry Letters*, *11*(3), 945–951. https://doi.org/10.1021/acs.jpclett.9b03154.

Jack, A., & Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy and R factor. *Acta Crystallographica. Section A, Foundations of Crystallography*, *34*(6), 931–935.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620.

Jeffries, C. M., Graewert, M. A., Blanchet, C. E., Langley, D. B., Whitten, A. E., & Svergun, D. I. (2016). Preparing monodisperse macromolecular samples for successful biological small-angle X-ray and neutron-scattering experiments. *Nature Protocols*, *11*(11), 2122–2153.

Jussupow, A., Messias, A. C., Stehle, R., Geerlof, A., Solbak, S. M., Paissoni, C., et al. (2020). The dynamics of linear polyubiquitin. *Science Advances*, *6*(42), eabc3786.

Kimanius, D., Pettersson, I., Schluckebier, G., Lindahl, E., & Andersson, M. (2015). SAXS-guided metadynamics. *Journal of Chemical Theory and Computation*, *11*(7), 3491–3498.

Knight, C. J., & Hub, J. S. (2015). WAXSiS: A web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nucleic Acids Research*, *43*, W225–W230.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Manalastas-Cantos, K., Konarev, P. V., Hajizadeh, N. R., Kikhney, A. G., Petoukhov, M. V., Molodenskiy, D. S., et al. (2021). *ATSAS 3.0*: Expanded functionality and new tools for small-angle scattering data analysis. *Journal of Applied Crystallography*, *54*(1), 343–355. https://doi.org/10.1107/S1600576720013412.

Merzel, F., & Smith, J. C. (2002). Is the first hydration shell of lysozyme of higher density than bulk water? *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 5378–5383.

Moore, P. B. (1980). Small-angle scattering. Information content and error analysis. *Journal of Applied Crystallography*, *13*(2), 168–175.

Moore, P. B. (2014). The effects of thermal disorder on the solution-scattering profiles of macromolecules. *Biophysical Journal*, *106*, 1489–1496.

Paissoni, C., Jussupow, A., & Camilloni, C. (2020). Determination of protein structural ensembles by hybrid-resolution SAXS restrained molecular dynamics. *Journal of Chemical Theory and Computation*, *16*(4), 2825–2834. https://doi.org/10.1021/acs.jctc.9b01181.

Park, S., Bardhan, J. P., Roux, B., & Makowski, L. (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.

Pelikan, M., Hura, G. L., & Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle X-ray scattering. *General Physiology and Biophysics*, *28*(2), 174.

Petoukhov, M. V., & Svergun, D. I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. *Biophysical Journal*, *89*(2), 1237–1250.

Pitera, J. W., & Chodera, J. D. (2012). On the use of experimental observations to bias simulated ensembles. *Journal of Chemical Theory and Computation*, *8*(10), 3445–3451. https://doi.org/10.1021/ct300112v.

Putnam, C. D., Hammel, M., Hura, G. L., & Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: Defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly Reviews of Biophysics*, *40*(3), 191–285. https://doi.org/10.1017/S0033583507004635.

Rambo, R. P., & Tainer, J. A. (2013). Accurate assessment of mass, models and resolution by small-angle scattering. *Nature*, *496*(7446), 477.

Ravera, E., Sgheri, L., Parigi, G., & Luchinat, C. (2016). A critical assessment of methods to recover information from averaged data. *Physical Chemistry Chemical Physics*, *18*(8), 5686–5701.

Rieping, W., Habeck, M., & Nilges, M. (2005). Inferential structure determination. *Science*, *309*(5732), 303–306.

Rout, M. P., & Sali, A. (2019). Principles for integrative structural biology studies. *Cell*, *177*(6), 1384–1403. https://doi.org/10.1016/j.cell.2019.05.016.

Roux, B., & Weare, J. (2013). On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *The Journal of Chemical Physics*, *138*(8), 084107.

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

Shevchuk, R., & Hub, J. S. (2017). Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. *PLOS Computational Biology*, *13*, e1005800.

Tiede, D. M., Zhang, R., & Seifert, S. (2002). Protein conformations explored by difference high-angle solution x-ray scattering: Oxidation state and temperature dependent changes in cytochrome c. *Biochemistry*, *41*(21), 6605–6614.