PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics

Roman Shevchuk[1,2], **Jochen S. Hub**[1,2]*

1 Institute for Microbiology and Genetics, University of Göttingen, Göttingen, Germany, **2** Göttingen Center for Molecular Biosciences (GZMB), University of Goettingen, Goettingen, Germany

* jhub@gwdg.de

OPEN ACCESS

Citation: Shevchuk R, Hub JS (2017) Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. PLoS Comput Biol 13(10): e1005800. https://doi.org/10.1371/journal.pcbi.1005800

Editor: Anders Wallqvist, US Army Medical Research and Materiel Command, UNITED STATES

Received: April 6, 2017

Accepted: September 29, 2017

Published: October 18, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by the Deutsche Forschungsgemeinschaft (grant No. HU 1971/1-1, HU 1971/3-1, HU 1971/4-1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Small-angle X-ray scattering is an increasingly popular technique used to detect protein structures and ensembles in solution. However, the refinement of structures and ensembles against SAXS data is often ambiguous due to the low information content of SAXS data, unknown systematic errors, and unknown scattering contributions from the solvent. We offer a solution to such problems by combining Bayesian inference with all-atom molecular dynamics simulations and explicit-solvent SAXS calculations. The Bayesian formulation correctly weights the SAXS data versus prior physical knowledge, it quantifies the precision or ambiguity of fitted structures and ensembles, and it accounts for unknown systematic errors due to poor buffer matching. The method further provides a probabilistic criterion for identifying the number of states required to explain the SAXS data. The method is validated by refining ensembles of a periplasmic binding protein against calculated SAXS curves. Subsequently, we derive the solution ensembles of the eukaryotic chaperone heat shock protein 90 (Hsp90) against experimental SAXS data. We find that the SAXS data of the apo state of Hsp90 is compatible with a single wide-open conformation, whereas the SAXS data of Hsp90 bound to ATP or to an ATP-analogue strongly suggest heterogenous ensembles of a closed and a wide-open state.

Author summary

In solution, many proteins adopt ensembles of multiple distinct states. The relative concentrations of the states are tightly controlled by factors such as pH, phosphorylation, or ligand binding, and a misbalance between the states underlies diseases such as cancer or neurodegeneration. However, detecting protein ensembles in experimental data has remained challenging. We present a statistically founded procedure for refining protein structures and ensembles against X-ray solution scattering data by combining atomistic simulations with Bayesian inference.

Introduction

Proteins are dynamic nanomachines that often populate heterogeneous ensembles of multiple distinct structural states. Controlling the relative population of such states is pivotal for the correct functioning of biological cells, and any misbalance between states may lead to severe conditions such as cancer or neurodegeneration. Detecting, understanding, and manipulating heterogeneous protein ensembles has therefore remained a central goal of molecular biophysics [1].

Deriving solution ensembles of proteins from structural experimental data has remained challenging, mainly because the information content of the data is typically insufficient to define all degrees of freedom of the ensemble [2, 3]. Consequently, upon fitting of structures or ensembles against experimental data, the data must be complemented by a physical model that restrains the protein into physically reasonable conformations, thereby reducing the risk of overfitting the model. Bayesian inference provides a route founded on probability theory for combining experimental data with physical models [4]. Applied to structure determination, Bayesian inference may become computationally expensive and technically challenging since it requires explicit sampling of the conformational space of the protein. However, it also holds a number of key advances over more simple optimization algorithms, as it provides statistically founded procedures (i) to weight the experimental data versus prior physical knowledge, and (ii) to quantify the uncertainty (or ambiguity) of the fitted structural model [5]. Due to its probabilistic rigor, Bayesian inference has been gaining increased popularity in various fields of biophysics, and it has hence been successfully applied for the refinement of structures against restraints from NMR, EPR, cryo-EM, and single-particle X-ray diffraction [5–11]. Following the pioneering work by Rieping et al., we refer to structural modeling based on Bayesian statistics as 'inferential structure determination' (ISD) [6].

Small-angle X-ray scattering (SAXS) is an increasing popular method that is in principle capable of detecting biomolecular structures and ensembles in solution [12, 13]. However, due to the low information content of SAXS data, refining structures or ensembles without overfitting poses a major challenge. For the refinement of individual structures against SAXS data, two routes have been suggested to reduce the risk of overfitting: first, during refinement, nearly all degrees of freedom of the biomolecule are constrained, leading to methods such as rigid-body modelling or normal mode fitting [14–17]. Second, physical information may be added to the low-information SAXS data, for instance by coupling a force field-based molecular dynamics (MD) simulation to the data with an energetic restraint [18–20]. Here, we follow the second route, building upon our method of SAXS-driven MD simulations [18]. SAXS-driven MD simulations drive biomolecular structures into conformations that are compatible with the data, using a differentiable harmonic restraint to the data. Critically, the method employs explicit-solvent calculations for predicting SAXS curves from the simulations frames, which were shown to provide accurate prediction for small and wide angles without the need of adjusting fitting parameters for the hydration layer or excluded solvent (see Fig 1) [21, 22]. In other words, the method uses a highly accurate and predictive 'forward model'. However, as formulated previously, the method was not Bayesian and, consequently, did not yet benefit from advantages of ISD-related approaches (see above).

Many methods for the refinement of heterogenous ensembles against experimental data follow a "sample-and-select" strategy [23]. Accordingly, first, an ensemble is proposed by sampling from a computationally efficient physical model, such as a coarse-grained force field. Second, a limited number of structures or clusters are picked from the proposed ensemble. Third, the weights of the structures or clusters are modified in a statistically meaningful manner until the data back-calculated from the refined ensemble agrees with the given experimental

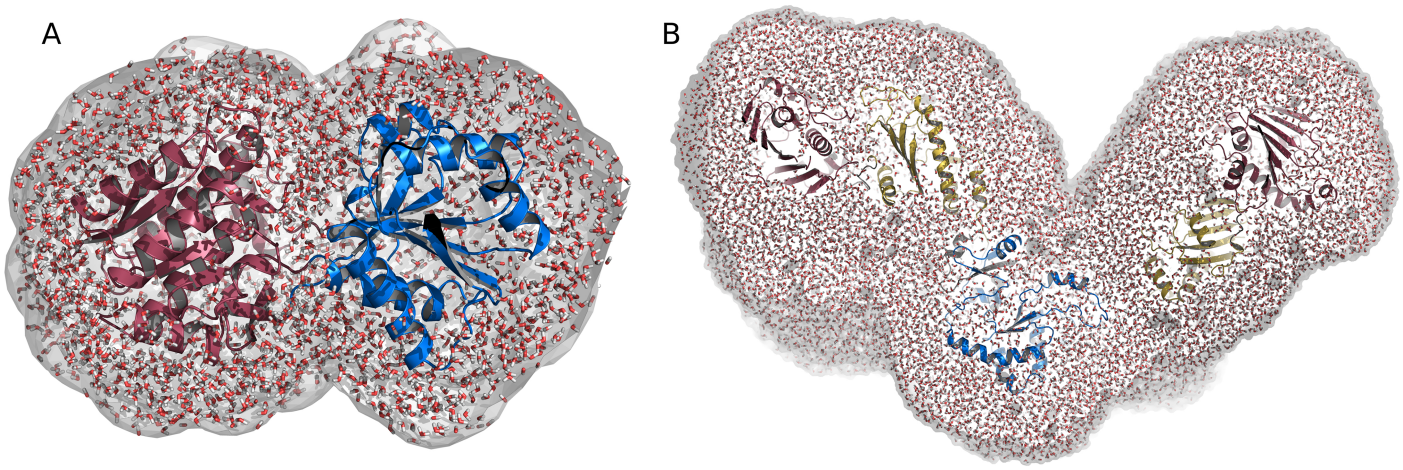

Fig 1. Envelopes of (A) leucine binding protein (LBP) and (B) heat shock protein 90 (Hsp90), illustrated as a white surface. Water molecules (red/white sticks) inside the envelope contributed to the explicit-solvent calculations used to compute the SAXS curves and the SAXS curve gradients, as required for the refinement simulations. (A) The N- and C-terminal domains of LBP are shown in red and blue cartoon representation, respectively. (B) The N-terminal, middle, and C-terminal domains of Hsp90 are shown in red, yellow, and blue cartoon representation, respectively.

https://doi.org/10.1371/journal.pcbi.1005800.g001

data. Examples of this strategy in the context SAXS are the EROS, BSS-SAXS, and EOM methods [24–26], yet a number of related approaches have been suggested ([23] and references therein). Interesting recent developments proposed the Bayesian derivation of continuous ensembles from experimental data, including a development for SAXS data [27, 28].

In this work, we take an alternative approach for the refinement of ensembles against SAXS data. First, following the ISD approach, we embed SAXS-driven MD simulations into a Bayesian inference framework. Hence, we derive *posterior* distributions of protein structures in the light of the SAXS data and the applied force field. Because we simulate with a physically accurate all-atom force field, we employ an accurate and informative *prior* of protein structures. Additional unknown parameters, termed nuisance parameters in context of ISD [6], are not chosen ad hoc but instead estimated simultaneously with the protein structures. Specifically, two remaining fitting parameters as well as a systematic uncertainty due to the buffer subtraction are taken as nuisance parameters. Second, we extend the concept of ISD towards an ensemble of a small number of structural states, allowing us to estimate the structural weights simultaneously with the structures and the nuisance parameters. In combination, due to the commitment to Bayesian inference, the new method provides statistically founded confidence intervals for both the structures *and* the structural weights. In addition, we show that the posterior distribution for the structural weights can be used as a criterion for detecting the number of states in the ensemble that is required to explain the data.

Bayesian interpretation of SAXS data: Goal of the method

We consider proteins that adopt an ensemble of a small number of distinct states. Typical examples would be proteins that exist in a mixture of active and inactive states, apo and holo states, or in a mixture of a few states along a more complex conformational cycle. We aim to derive the coordinates $\mathbf{R} = (\mathbf{R}_1, \ldots, \mathbf{R}_N)$ as well as the relative weights (or concentrations) $\mathbf{w} = (w_1, \ldots, w_N)$ of the states from given experimental SAXS data, where N is the number of states. Hence, the term 'ensemble' does not refer to the thermodynamic ensemble, but instead to a specific set $(\mathbf{R}, \mathbf{w})$. Notably, since the ensemble reduces to a single structure by setting $N = 1$,

Bayesian structure refinement (instead of ensemble refinement) is contained in the method presented here as a special case.

Posterior distribution of the ensemble

Since the number of independent data points in a SAXS curve is much smaller than the number of degrees of freedom of any protein, it is unlikely that only a single ensemble (**R**, **w**) fits the SAXS data, but instead a wide range of ensembles are typically compatible with the data. A statistically founded procedure to infer the ensemble from the data can be formulated as the conditional probability $p(\mathbf{R}, \mathbf{w}, \theta|D, K)$ that quantifies the plausibility of the ensemble (**R**, **w**) in the light of the SAXS data D and prior physical knowledge K [4]. The symbol θ summarizes nuisance parameters, which are of limited interest but required for evaluating whether the ensemble (**R**, **w**) is compatible with the data D (see below). The posterior distribution is most conveniently evaluated using Bayes' theorem:

$$p(\mathbf{R}, \mathbf{w}, \theta|D, K) \propto L(D|\mathbf{R}, \mathbf{w}, \theta, K)\pi(\mathbf{R}|K)\pi(\mathbf{w}|K)\pi(\theta|K). \quad (1)$$

Here, $L(D|\mathbf{R}, \mathbf{w}, \theta, K)$ denotes the likelihood that the data D were measured given the ensemble (**R**, **w**) and nuisance parameters θ. The functions $\pi(\mathbf{R}|K)$, $\pi(\mathbf{w}|K)$, and $\pi(\theta|K)$ denote the prior distributions of possible protein conformations, weights, and nuisance parameters. Due to the low information content of SAXS data, $L(D|\mathbf{R}, \mathbf{w}, \theta, K)$ provides only limited information, i.e., L is a wide function of R. Hence, in order to draw structural conclusions from the data, that is, to arrive at a reasonably narrow posterior distribution, it is critical to impose an informative tight prior $\pi(\mathbf{R}|K)$ of protein conformations, which is here achieved by applying an accurate physical model. In the method presented here, the prior $\pi(\mathbf{R}|K)$ is naturally given through an unbiased MD simulation, where K represents the physical laws and the force field underlying the MD simulation.

Accounting for unknown systematic errors

Formulating a likelihood function L for SAXS refinement is not straightforward because experimental SAXS data report purely statistical errors, whereas systematic errors, for instance due to poor buffer matching, are typically unknown. For data recorded at modern single-photon counting detectors, systematic error may dominate the overall uncertainty, suggesting that systematic errors strongly contribute to the likelihood L. In addition, for comparing experimental with calculated SAXS curves, free fitting parameters must be adjusted [21, 29–31]. Since both the systematic errors and fitting parameters are a priori unknown, a full Bayesian treatment requires that those parameter are simultaneously estimated with the structures and weights. Hence, systematic errors as well as fitting parameters are treated as nuisance parameters θ in the present method.

In practice, one is mainly interested in the ensemble (**R**, **w**), but not in the nuisance parameters θ. The statistically correct way of reducing the general posterior p in Eq 1 to the posterior of the structural ensemble is to marginalize out the nuisance parameters,

$$p(\mathbf{R}, \mathbf{w}|D, K) = \int \mathrm{d}\theta\, p(\mathbf{R}, \mathbf{w}, \theta|D, K). \quad (2)$$

In our method, the fitting parameters are marginalized out analytically at the level of the likelihood, whereas systematic errors are explicitly sampled and marginalized out numerically (Methods and materials). To visualize the high-dimensional $p(\mathbf{R}, \mathbf{w}|D, K)$, the posterior may be further projected onto intuitively important coordinates, such as the distance between two protein domains, the radius of gyration of the protein, or the weight of an interesting state.

Energy analogs

By taking the negative logarithm of the posterior, Eq 1 takes the form of a hybrid energy that is commonly applied for structure refinement [5, 9], yet corrected with the contributions from the prior distributions:

$$E_{\text{hybrid}} = V_{\text{ff}}(\mathbf{R}, K) + E_{\text{exp}}(\mathbf{R}, \mathbf{w}, \theta, D, K) - \beta^{-1} \ln [\pi(\mathbf{w}|K)\, \pi(\theta|K)]. \quad (3)$$

Here, the posterior was identified with a hybrid energy $E_{\text{hybrid}} = -\beta^{-1} \ln p(\mathbf{R}, \mathbf{w}, \theta|D, K)$, where β denotes the inverse temperature. The prior for the protein structures is taken from the MD force field energy as $V_{\text{ff}}(\mathbf{R}, K) = -\beta^{-1} \ln \pi(\mathbf{R}|K)$, after marginalizing out the solvent coordinates (see Methods and materials). The experiment-derived energy is given via the likelihood, $E_{\text{exp}} = -\beta^{-1} \ln L$, adding an energetic penalty if the SAXS curve calculated from the ensemble $(\mathbf{R}, \mathbf{w})$ is incompatible with the data D.

Sampling the posterior distribution

Having translated the probabilities into energies, all parameters can be sampled using established methods. Accordingly, sampling of protein structures $\mathbf{R}$ is conducted using Newtonian dynamics. Here, the force on atom ℓ is given via gradients of the hybrid energy with respect to the atomic positions, $\mathbf{F}_\ell = -\nabla_\ell E_{\text{hybrid}}$, evaluated at fixed $\mathbf{w}$ and fixed nuisance parameters. The fitting parameters, as shown below, are marginalized analytically at the level of the likelihood. The remaining nuisance parameter, namely the systematic error σ_{buf}, as well as the weights $\mathbf{w}$ are sampled using Gibbs sampling, that is, Monte-Carlo moves at fixed protein coordinates $\mathbf{R}$. Calculations of the SAXS intensity and intensity gradients from $(\mathbf{R}, \mathbf{w})$, as required for sampling the posterior, were conducted with the explicit-solvent algorithms established previously [18, 21], taking accurate atomistic models for both the hydration layer and the excluded solvent (Fig 1). Details on the likelihood function, assumed priors, force calculations, and sampling algorithms are provided in the Methods and Materials.

A probabilistic criterion for choosing the number of states

The weights $\mathbf{w}$ are normalized and have non-negative elements, i.e., the relevant weight space is given by the $(N-1)$-simplex. Sampling of the weight space was accelerated using umbrella sampling along the weights [32]. This is computationally convenient because it allows calculation of the posterior from a set of short independent simulations. More critically, this allows us to compute the posterior for the complete weight space, including the "edge" of the simplex, where at least one of the weights w_j is zero $(1 \leq j \leq N)$. However, note that weight vectors $\mathbf{w}$ with elements equal to zero specify smaller ensembles with a reduced number of states. Consequently, the posterior of an ensemble of N states includes all smaller ensembles as a special case, thereby proving a probabilistic criterion for choosing the number of states required to explain the experimental data: if the posterior peaks at the edge of the simplex, a smaller ensemble provides a plausible model; in turn, if the posterior near the edge is small compared to the posterior's maximum, a smaller ensemble is implausible.

Results

In the following, Bayesian ensemble refinement is demonstrated for two test proteins: leucine binding protein (LBP) using calculated SAXS data and heat shock protein 90 (Hsp90) using experimental SAXS data. We assumed that both proteins adopt a two-state ensemble of an open and a closed structure $(N = 2)$. We further assume that the closed structure is known, whereas (i) the coordinates of the open structure as well as (ii) the relative open/closed weights

are simultaneously refined against SAXS data. Such scenarios are quite common, as a compact holo or ground state structure might be accessible to X-ray crystallography, whereas more flexible apo or excited state structures often do not crystallize. Applying the method proposed here to larger ensembles of $N > 2$ is conceptually possible but beyond the scope of this article. With increasing number of states N, due to increasing number of required umbrella windows, the computational cost would scale exponentially with $N - 1$.

Leucine binding protein (LBP)

LBP is a typical representative of the superfamily of periplasmic binding proteins involved in chemotaxis and solute uptake over membranes [33]. LBP is a well-characterized two-domain protein, exhibiting a transition from an open (apo) to a closed (holo) state triggered by ligand binding (Fig 2A/2B) [34, 35]. Free simulations of the closed and open state suggested center-of-mass distances between the N- and C-terminal domains of ∼3 nm and 3.25 nm, respectively, which is compatible with experimental SAXS data of the homologous LIVBP [18]. We theoretically computed SAXS curves of the open and closed states (Fig 2C, solid lines), as well as linear combinations, thereby modeling SAXS data from heterogeneous ensembles of known open/closed weights of 0:100, 25:75, 50:50, 75:25, and 100:0 (Fig 2C, dashed lines).

The posteriors of the ensembles $p(\mathbf{R}, \mathbf{w}|D, K)$ refined against these five SAXS curves are presented in Fig 2D. To visualize the high-dimensional posterior, we projected the posterior onto two characteristic coordinates: (i) the weight of the open state w_{open}, implying the weight $(1 - w_{\text{open}})$ for the closed state, and (ii) the interdomain distance d_{NC} of the open state (illustrated in Fig 2B). Evidently, all derived posterior distributions peak at the correct w_{open}. In addition, the posteriors refined against SAXS curves of non-zero open-state content (Fig 2D, four right panels) peak at the physically correct interdomain distance of ∼3.25 nm (Fig 2, see also the marginalized posteriors in S1A Fig). In addition, the RMSD to the mean open structure taken from umbrella simulations, restrained to weights at the maxima of the respective posterior, reveals that the refinement simulations rapidly approach the correct open state (S5 Fig). These findings demonstrate that the MD simulations were capable of translating the information in the SAXS curve into the underlying heterogeneous open/closed ensemble.

The width of the posteriors rigorously quantify the degree of structural knowledge that can (and cannot) be inferred from the SAXS curve, i.e., the posteriors quantify the ambiguity of the refined ensemble. For the LBP ensemble refinement, the marginalized posteriors suggest 65% confidence intervals (CI) for w_{open} and d_{NC} in the order of ±15% and ±0.07 nm, respectively (Fig 3A and S1A Fig, S1 and S2 Tables). In addition, the posteriors in Fig 2D suggest some correlation between w_{open} and d_{NC}, as apparent from the posterior's diagonal elongated shapes, suggesting that the SAXS curves are compatible with an increased w_{open} given that the open state is modeled more compact.

Single-state or two-state ensemble? Fig 3A presents the posterior distributions $p(w_{\text{open}}|D, K)$, as derived from the two-dimensional posteriors $p(w_{\text{open}}, d_{\text{NC}}|D, K)$ (Fig 2D) by marginalizing out the interdomain distance d_{NC}. As described in the Theory section, the odds that a single state versus a two-state ensemble underlie the SAXS data is quantified by $p(w_{\text{open}}|D, K)$ at the "edge" of the weight space, at $w_{\text{open}} = 0$ and $w_{\text{open}} = 1$, as compared to $p(w_{\text{open}}|D, K)$ at intermediate w_{open}. Fig 3B compares the posterior maximum at the edge p_{edge} with the posterior maximum in the entire weights space p_{max}, plotted as $p_{\text{edge}}/p_{\text{max}}$. Evidently, the posteriors refined against SAXS curves representing purely the closed or purely the open state exhibit a peak at the edge, thus recovering that a single state is sufficient to explain this data (Fig 3A/3B, 0:100 and 100:0). In contrast, posteriors refined against SAXS curves of heterogeneous ensembles are small at the edge and instead peak at intermediate w_{open} (Fig 3A/3B, 25:75, 50:50,

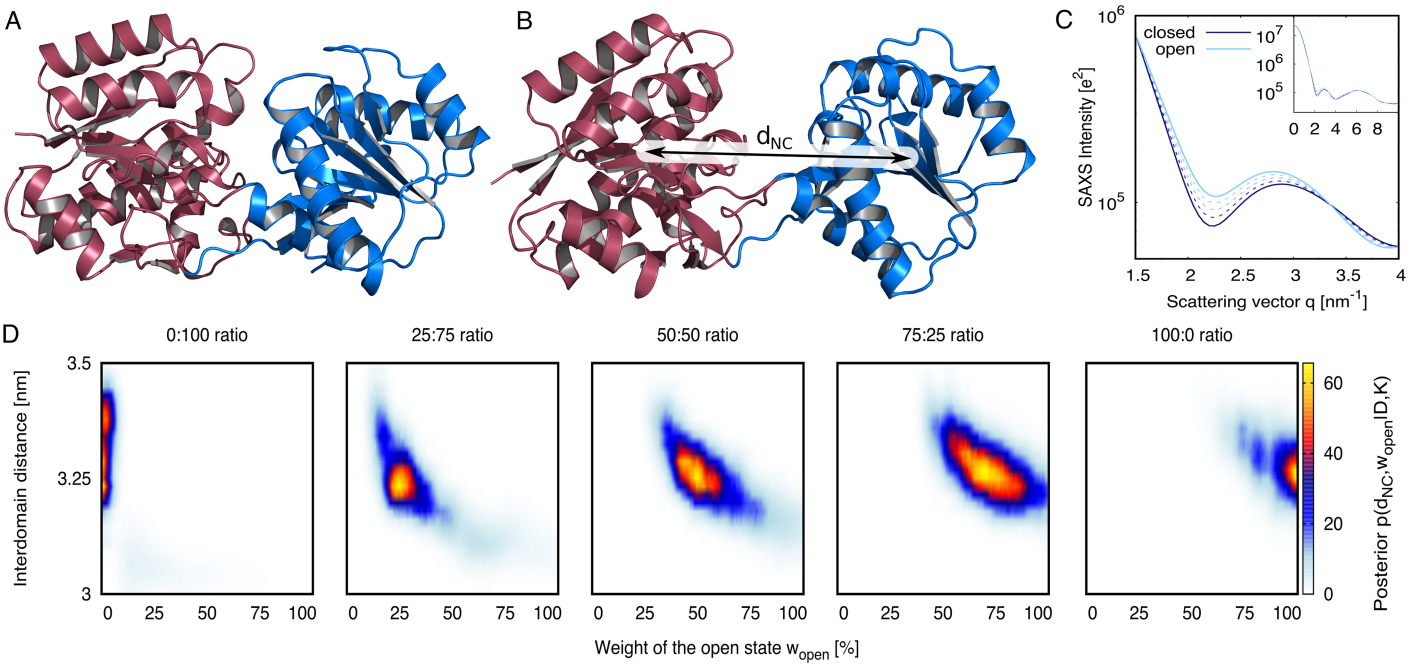

Fig 2. Bayesian ensemble refinement of leucine binding protein (LBP). (A) Cartoon representation of LBP in closed and (B) open states. N- and C-terminal domains are colored in red and blue, respectively. (C) Solid lines: Computed SAXS curves of the open (light blue) and closed state (dark blue). Dashed lines: SAXS curves of open/closed heterogeneous ensembles, computed with open/closed weights of 25:75, 50:50, and 75:25. Inset: Complete SAXS curves up to $q = 10$ nm^{-1}. Large figure: closeup view highlighting the differences between the SAXS curves. (D) Posterior distribution of the refined two-state ensemble, projected onto the weight w_{open} and the interdomain distance d_{NC} of the open state. The one-dimensional marginalized distributions are shown in Fig 3A and S1 Fig.

https://doi.org/10.1371/journal.pcbi.1005800.g002

75:25). Hence, our method recovers that a single state is highly implausible in the light of these SAXS curves and the force field, and that instead two states are required to explain the data.

It is instructive to compare the two-state refinement presented in Figs 2 and 3 with an attempt to interpret the five SAXS curves of Fig 2D by a single state. To this end, we refined a single state ($N = 1$) against each of the five computed SAXS curves. As expected, refining a

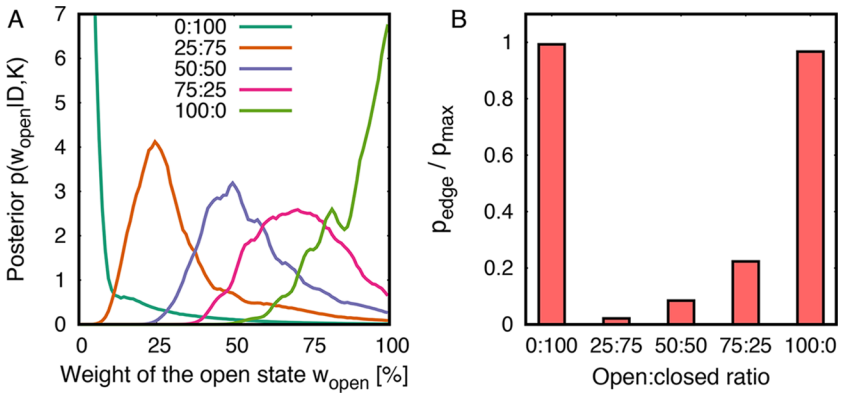

Fig 3. (A) Marginalized posterior of the weight of the open state from refinement of leucine binding protein (LBP). (B) Odds that a single state vs. a two-state ensemble underlies the SAXS curves, presented as p_{edge}/p_{max}. For details, see text.

https://doi.org/10.1371/journal.pcbi.1005800.g003

single state against the SAXS curves of purely the closed or purely the open state (100:0 or 0:100) recovers the correct interdomain distances of ∼3.0 nm and ∼3.25 nm for the closed and open states, respectively (S1B Fig, S3 Table). In contrast, refining a single state against SAXS curves that, in truth, represent a heterogeneous open/closed ensemble (25:75, 50:50, 75:25) leads to a misinterpretation of the SAXS data in terms of intermediate partially open states (S1B Fig, S3 Table), contrasting the fact that such intermediate states are hardly populated in a free microsecond-long simulation. Critically, the fitted SAXS curves well match the target curves, suggesting that a visual inspection of the fitted curves is insufficient to reveal that such partially open states are a misinterpretation (S2 Fig). Hence, an analysis similar to the Bayesian inference on the number of states, as presented in Fig 3, is indeed required to detect the correct number of states from the SAXS curve.

Heat shock protein 90

Hsp90 is a chaperone that interacts with more than 200 proteins in eucaryotic cells [39–42]. It constitutes up to 2% of the cellular protein mass [43]. Since many client proteins of Hsp90 are oncogenic, Hsp90 has been suggested as a promising target for anti-cancer therapies [44, 45]. Structurally, Hsp90 is a homodimer, where each protomer contains three domains: an N-terminal domain with the ATP binding site, a middle domain forming the interaction sites for client proteins, and a C-terminal domain responsible for Hsp90 dimerization (Fig 4A). Crystallographic, cryo-EM, and SAXS studies established that Hsp90 carries out large-scale conformational transitions between a V-shaped open state and a compact closed state, controlled by binding of ATP, ATP analogues, and client proteins [38, 46]. However, ligands do not dictate a single well-defined state, but instead merely shift the equilibria of heterogeneous ensembles between open and closed conformations [47, 48]. Only recently it was found that sufficient time spent in the open state is crucial for correct Hsp90 functioning, highlighting the importance of controlling the open/closed equilibria of the chaperone [49].

Based on experimental SAXS data of yeast Hsp90 in the apo state, Hsp90 bound to the slowly hydrolyzing ATP-analogue AMPPNP, and Hsp90 bound to ATP (Fig 4C, colored curves [37]), we derived heterogeneous solution ensembles of the Hsp90 dimer using Bayesian ensemble refinement. Hsp90 ensembles were modeled as two-state ensemble of (i) the closed state, taken from the yeast crystal structure (Fig 4A), and (ii) an initially unknown open state. Starting the simulations from a nearly closed conformation, both the structure and the relative weight w_{open} of the open state were simultaneously refined against the SAXS data. The SAXS curves of the refined two-state ensembles exhibited reasonable agreement with the experimental curves (Fig 4C).

The residuals between calculated and experimental SAXS curves are analyzed in Fig 5. Here, panel (A) shows the residuals normalized purely by the statistical experimental errors, $\Delta I(q)/\sigma_{\mathrm{exp}}$. The large residuals at low q (Fig 5A, red and yellow) reflect that the MD force field did not allow conformations that accurately fullfil the data within statistical experimental errors, possibly because accurately reproducing the data would require an energetically unfavorable conformational transition (such as partial unfolding). In other words, the Bayesian analysis revealed that, in the light of the force field, substantial systematic errors at low q are more plausible than an ensemble that accurately matches the experimental data. Indeed, as shown in Fig 5B the residuals normalized with respect to the total errors including statistical *and* systematic errors, $\Delta I(q)/\sigma_{\mathrm{tot}}$, reveal reasonably low values over the entire q-range. As outlined in the Methods, we modelled systematic errors as a consequence of poor buffer matching, but the analysis can not exclude other sources of remaining discrepancies such as a small fraction of aggregated Hsp90. Further, in this work, we can not fully exclude the possibility

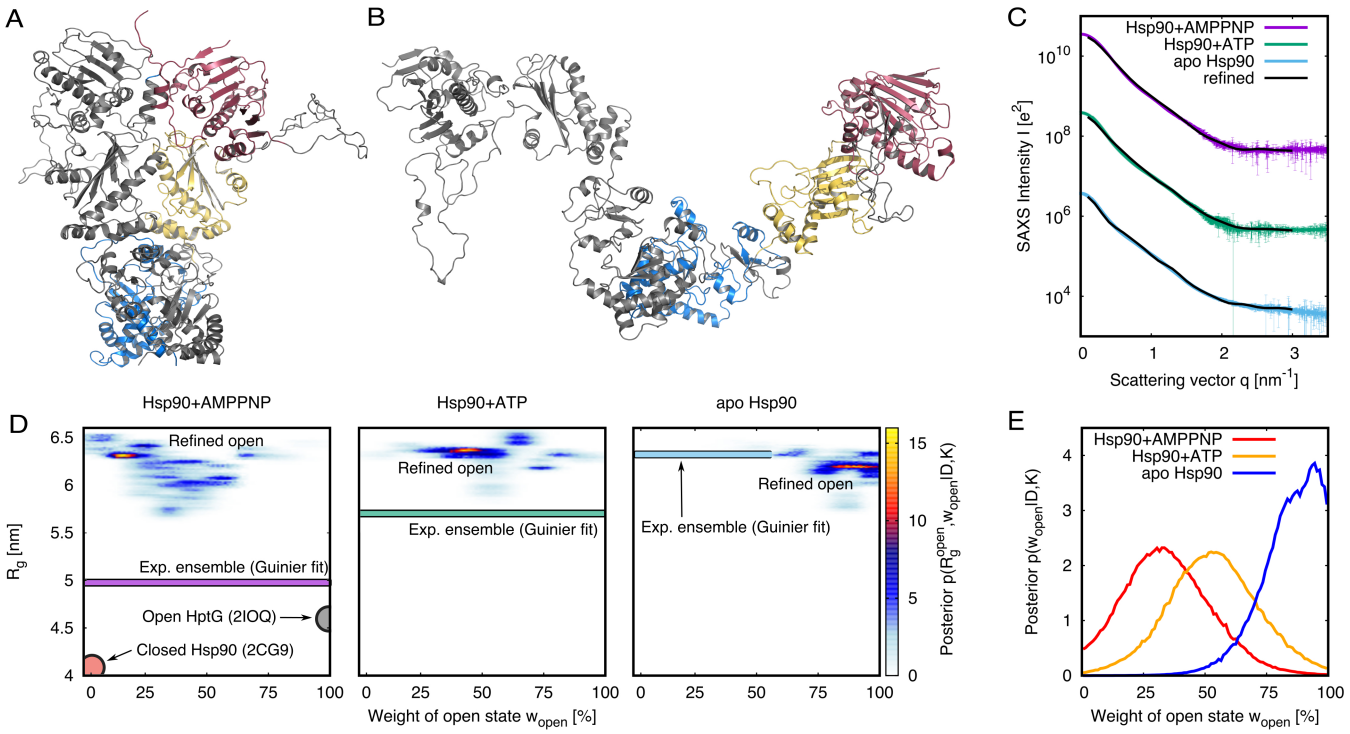

Fig 4. Bayesian ensemble refinement of Hsp90. (A/B) Cartoon representation of Hsp90, the first protomer shown in grey, and the second protomer in color. C-terminal, middle, and N-terminal domains are shown in blue, yellow, and red, respectively. (A) Closed state, modelled from the 2CG9 structure [36]. (B) Open state, refined against SAXS data. (C) Experimental SAXS curves (colored lines, taken from ref. [37]) and calculated SAXS curves (black) computed from the refined ensembles. For clarity, curves for AMPPNP- and ATP-bound states were vertically offset, and experimental data points with very large errors were removed. (D) Posterior distribution of refined Hsp90 ensemble plotted as function of the weight (w_{open}) and radius of gyration (R_g^{open}) of the open state. For reference, The radius of gyration R_g of the closed Hsp90 structure (2CG9) and of the partially open *E. coli* HptG structure (2IOQ) are indicated as pale red and grey circles [36, 38]. In addition, R_g of the experimental ensembles, taken from Guinier fits to the SAXS curves in (C), are indicated as colored bars (color coding according to C). (E) Posterior of w_{open}, computed from the maps in (D) by marginalizing out R_g.

https://doi.org/10.1371/journal.pcbi.1005800.g004

that a more continuous ensemble, as supported by recent Förster resonance energy transfer (FRET) study [50], might provide a more accurate description of the experimental conditions.

Fig 4D presents the posterior distributions $p(\mathbf{R}, \mathbf{w}|D, K)$ of the Hsp90 ensembles, projected onto two intuitive degrees of freedom: (i) the weight w_{open} of the open state, implying the weight $(1 - w_{\text{open}})$ for the closed state, and (ii) the radius of gyration R_g^{open} of the refined open state, which naturally quantifies the degree of opening of the open state. The marginal posteriors $p(R_g^{\text{open}}|D, K)$ for the three ensembles, obtained by marginalizing the posteriors in Fig 4D with respect to w_{open}, are presented in Fig 6 as colored lines. Evidently, the refined structures of the open state were similar in all three ensembles, exhibiting large R_g^{open} values of $\sim$6.3nm. These R_g^{open} values are $\sim$1.3 nm and $\sim$1.7 nm larger than the radii of gyration of the open form of the bacterial HtpG homologue in the crystal and in solution environment, respectively [38, 48], but they are compatible with previously reported open conformations of eukaryotic apo Hsp90s [51]. Hence, the open structures of the three refined open/closed heterogeneous ensembles are characterizing by a wide open conformation, as visualized in Fig 4B.

Fig 4E presents the marginal posteriors of the weight of the open state, $p(w_{\text{open}}|D, K)$, obtained by marginalizing the posteriors in Fig 4D with respect to R_g^{open}. Evidently, w_{open} strongly differs between the three ensembles. The posteriors suggest closed:open populations

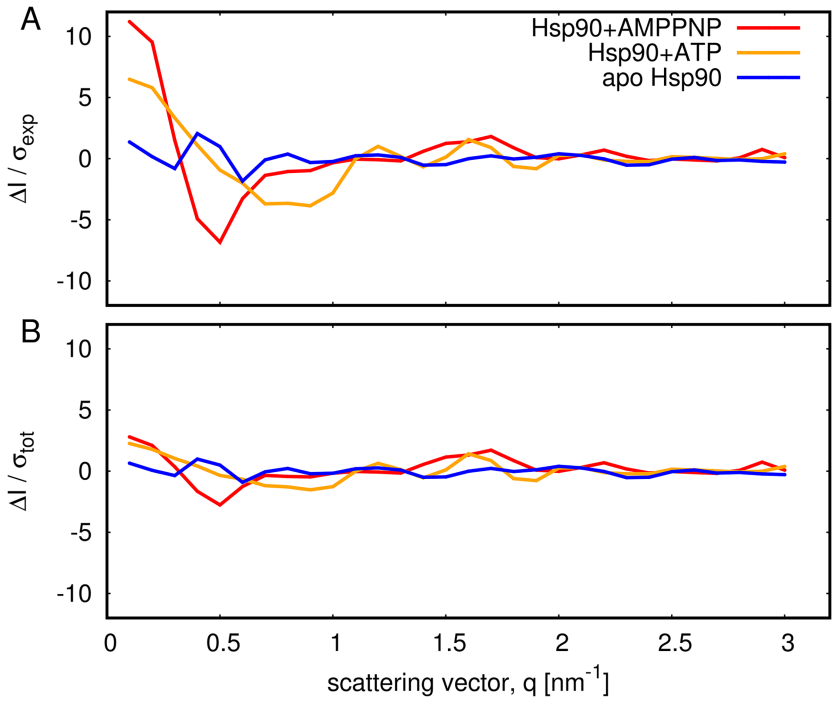

Fig 5. Residuals between the calculated SAXS curves and the experimental SAXS curves, evaluated at the *q*-points applied during the refinement simulations of Hsp90 (color code see legend). (A) Residuals $\Delta I/\sigma_{exp}$ normalized with respect to purely statistical experimental errors σ_{exp}. Large residuals at low q reflect that the MD force field prohibited structures that would accurately match the data within statistical errors. (B) Residuals $\Delta I(q)/\sigma_{tot}$, where σ_{tot} denotes the total error including both statistical and estimated systematic errors (see Methods for details). The reduced residuals compared to panel (A) reflect that the Bayesian analysis suggested substantial systematic errors as the most plausible explanation for discrepancies between calculated and experimental SAXS curves.

https://doi.org/10.1371/journal.pcbi.1005800.g005

of 68:32 and 52:48 for the AMPPNP- and ATP-bound states, respectively, with 65% confidence intervals of ±18% (S4 Table). Hence, for the AMPPNP- and ATP-bound states, a model of a single state is very implausible in the light of the MD force field and the SAXS data. These findings resemble results from rigid-body SAXS modeling of a bacterial HtpG homologue that suggested heterogeneous closed/open ensembles in the AMPPNP-bound state, yet without providing confidence intervals [48]. For the Hsp90 apo state, the posterior suggests that w_{open} is with 65% confidence within the interval [78%,100%], suggesting that a single open state as well as a heterogeneous ensemble with a large w_{open} are both compatible with the SAXS data and the MD force field.

Discussion

We have presented a method for the refinement of a single protein structure or of an ensemble of structures against SAXS data, applicable to ensembles of a small number of distinct states. By combining Bayesian inference with atomistic MD simulations, the method is capable of inferring the structures and structural weights that gave rise to the SAXS data. The method does not merely derive a single "best fit" against the experimental data, but instead provides the joint posterior distribution of structures and weights, thus quantifying the plausibility of all possible structures and ensembles in the light of data D and physical knowledge K. The width of the posteriors yield confidence intervals founded on probability theory for both the

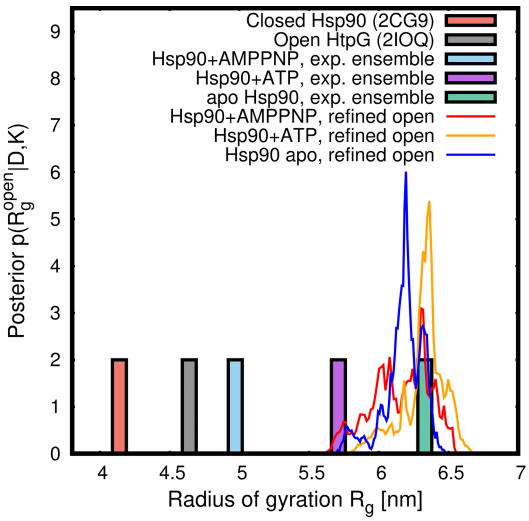

Fig 6. Lines: Marginalized posteriors of the radius of gyration R_g^{open} of the refined open state from two-state refinement of Hsp90+AMPPNP, Hsp90+ATP, and apo Hsp90 (for color code, see legend). The posteriors characterize a wide open conformation. For comparison, pale green, purple, and pale blue bars indicate the R_g values for the ensembles of Hsp90+AMPPNP, Hsp90+ATP, and apo Hsp90, respectively, estimated from a Guinier fit to the experimental data [37]. The pale red bar indicates R_g of the closed crystal structure of Hsp90 (2CG9), and the grey bar indicates R_g of the partially open crystal structure of *E. coli* HptG (2IOQ) [36, 38].

https://doi.org/10.1371/journal.pcbi.1005800.g006

structures and the weights, that is, the method quantifies the precision of the refined ensemble. Such reliable confidence intervals are required for deciding whether a structural model is convincingly supported by available SAXS data, or whether additional data are needed to unambiguously prove a model. We stress that the confidence intervals derived here should not be confused with the spread of "best fits" obtained by multiple repetitions of an optimization algorithm, as common, for instance, when fitting low-resolution bead models against SAXS data [52]. Repeated best fits test the convergence of the optimization algorithm but do not provide a statistically founded confidence interval.

Since we enforced exhaustive sampling of the weight space using umbrella sampling, the posterior includes smaller ensembles with a reduced number of states as a special case, as given by weight vectors **w** with one or multiple zero elements. We showed that this feature provides a rigorous criterion for deciding on the number of states required to explain the experimental data. For the apo state of Hsp90, we found that the SAXS data is compatible with a single open state, as well as with a heterogenous open-closed ensemble with a large weight of the open state. In contrast, for the AMPPNP- and ATP-bound states of Hsp90, we found that a single state is unlikely in the light of the SAXS data and the MD force field, whereas a model of two states provides a much more plausible model. Critically, Bayesian inference further allows us to assign a confidence to these qualitative statements. Namely, the odds that a single state underlies the SAXS data is 80% for the apo state, 20% AMPPNP-bound, and 6% for the ATP-bound state. As such, the researcher may decide whether such confidence is sufficient to decide on the number of states, or whether additional data, e.g. from FRET, should be included to further increase the confidence on the number of states [53].

A property of Bayesian methods is that the computed posterior depends on the chosen priors. For Bayesian SAXS refinement, the posterior $p(\mathbf{R}, \mathbf{w})$ most critically depends on the prior for protein conformations $\pi(\mathbf{R}, K)$, which is given through the applied force field. In this work,

we applied a physically accurate all-atom force field, which provides a more accurate description of the energy landscape as compared to rigid-body or coarse-grained force fields. However, despite major force field improvements in recent years [54], it can not be excluded that certain force fields bias the refinement simulations towards unphysical states, in particular for proteins with large disordered regions [55]. Hence, we recommend to use the most recent and best validated force fields.

Depending on the size of the system and the inherent autocorrelation times, exhaustive sampling of the posterior may become challenging. Due to the use of umbrella sampling along w_{open}, we here observed rapid convergence of the marginalized posterior $p(w_{\mathrm{open}}|D, K)$, both for LBP and Hsp90 (S6 Fig). The 2D posterior $p(w_{\mathrm{open}}, d_{\mathrm{NC}}|D, K)$ for LBP seemed converged at moderate computational effort of 50 ns per umbrella window (Fig 2D), whereas the 2D posterior $p(w_{\mathrm{open}}, R_g^{\mathrm{open}}|D, K)$ for Hsp90 converged more slowly, as apparently from the somewhat scattered posteriors (Fig 4D). Hence, in future refinement simulations of very large systems such as the system of Hsp90, the sampling may benefit from additional enhanced sampling methods.

The computational cost of the simulations presented here are increased by only $\sim$15% as compared to standard MD simulations, suggesting that the calculations are well feasible on modern hardware. However, MD simulations are obviously more expensive than simplified methods such as rigid-body modelling.

The sampling of structural weights has some similarity with previous sample-and-select methods that reweighed a set of structures against SAXS data using, for instance, Bayesian or maximum-entropy criteria [16, 24–26]. However, at variance with previous methods, we refined the weights *simultaneously* with the structures, fitting parameters, and systematic errors. This difference is not a technical subtlety but is instead critical to estimate the correct uncertainty of the weights: In our method, by commitment to Bayesian inference, the uncertainty (or ignorance) about the structure, fitting parameters, and systematic errors are propagated into the uncertainty of the weights. In other words, when estimating the weights, and in contrast to previous methods, we do not assume any precise knowledge about the structure, fitting parameters, and systematic errors that, in truth, we do not have. This difference rationalizes why the uncertainties of fitted weights reported previously are much smaller than the uncertainties derived here via the full Bayesian treatment [24].

The refinement simulations presented here differ from previous methods for structure and ensemble refinement against SAXS data by a number of additional elements. First, since our refinement simulations are steered by the experimental SAXS data, the simulations are capable of sampling large-scale conformational transitions, which would not be sampled in an equilibrium simulation due to limited simulation time. An example is the open/close transition of Hsp90 that occurs on the time scale of many seconds at experimental conditions [49]. As such, our method does not strictly require the application of coarse-grained simulations [24, 25] or other simplified physical models [16, 26] to visit the relevant conformational states. Second, because we apply purely the MD force field and the SAXS data but no additional constraints, the refinement is not limited to rigid-body motions or normal modes, which were previously used to refine structures against SAXS data [14–17]. Hence, prior to the refinement simulations, our method does not require the ad-hoc definition of rigid bodies and flexible linkers, which may not be obvious. Third, in contrast to previous refinement methods, SAXS curves were computed using explicit-solvent algorithms, avoiding any solvent-related fitting parameters [21, 56].

In this study, we built upon the concept of "inferential structure determination" (ISD), which was originally formulated to model NMR data with a single structural state [5, 6]. In

short, we presented an ISD method for SAXS data using all-atom MD simulations. In addition, we extended the ISD concept towards the refinement of a small number of states (typically two states), but the method is not intended for the refinement of continuous and highly heterogeneous ensembles. Hence, our approach complements methods for the reweighing of continuous ensembles against experimental data, as required for modeling of intrinsically disordered proteins [27, 57, 58], and it further complements maximum-entropy-based methods for biasing ensembles with experimental data [59–61].

We developed the method with a focus on SAXS data, but the calculations may be readily complemented by other sources of structural information. For instance, the refinement may be additionally guided by multiple sets of small-angle neutron scattering (SANS) data, optionally measured at various D_2O contrasts and differently deuterated solutes. Similar to the SAXS-guided refinement, such SANS-guided refinement simulations will benefit from the fitting-free explicit-solvent scattering calculations applied here. Alternatively, the refinement simulations may be complemented by additional distance restraints from double electronelectron resonance (DEER) or FRET. Such future developments, complementing the method proposed here, may provide a route to MD-based Bayesian integrative modeling.

Materials and methods

Modeling of systematic errors

A common source of systematic errors in SAXS experiments is poor buffer matching. We therefore modeled the systematic errors σ_{buf} as a consequence of a buffer density mismatch $\delta\rho_{\text{buf}}$ between the protein solution and the pure buffer. Following previous work [18], $\delta\rho_{\text{buf}}$ can be translated into an uncertainty σ_{buf} of the calculated intensity $I_c(q)$, contributing to the likelihood function (see below).

Likelihood function

We recently found excellent agreement of SAXS curves predicted from explicit-solvent MD simulations with experimental curves, if the experimental curves $I_{\text{exp}}(q)$ were adjusted by only two fitting parameters following $I_{\text{fit}}(q) = fI_{\text{exp}}(q) + c$, where f denotes the fitted absolute scale and c denotes a fitted constant offset [21], and q is the momentum transfer. Hence, we take for the likelihood function

$$L(D|\mathbf{R},\mathbf{w},\theta,K) \propto \exp\left[-\frac{N_{\text{indep}}}{2N_q}\sum_{i=1}^{N_q}\frac{[I_c(q_i,\mathbf{R},\mathbf{w})-(fI_{\text{exp}}(q_i)+c)]^2}{f^2\sigma_{\text{exp}}^2(q_i)+\sigma_{\text{calc}}^2(q_i)+\sigma_{\text{buf}}^2(q_i;\delta\rho_{\text{buf}})}\right], \quad (4)$$

where $\theta = \{f, c, \delta\rho_{\text{buf}}\}$. As shown below, the fitting parameters f and c can be marginalized out analytically. The symbols σ_{exp} and σ_{calc} denote statistical errors of the experimental and calculated intensities, respectively. The calculated intensity I_c is a weighted average over the intensities of the N states, $I_c(q_i,\mathbf{R},\mathbf{w}) = \sum_{j=1}^{N} w_j I(q_i,\mathbf{R}_j)$. The symbols N_q and N_{indep} denote the total and the independent number of data points in the SAXS curve, respectively. N_{indep} was estimated by the number of Shannon-Nyqvist channels $N_{\text{indep}} = q_{\max} D_p/\pi$, where D_p is the maximum diameter of the protein and $q_{\max}$ is the maximum momentum transfer of the SAXS curve [13]. Hence, the factor N_{indep}/N_q is an empirical correction that accounts for the fact that the number of independent data points N_{indep} in a SAXS curve is typically much smaller than the number of q-points N_q reported in experimental SAXS curves. Without the factor N_{indep}/N_q, the information content in the data would be overrated in comparison with the information in the priors. Critically, this correction assumes that the data I_{exp} corresponds to a

"smoothed" SAXS curve, and that the experimental errors $\sigma_{\mathrm{exp}}(q_i)$ denote the true uncertainty of point q_i in the light of correlations of I_{exp} along q.

Prior distributions

A flat prior was applied for the fitting parameters, $\pi(f) = \pi(c) = 1$. Notably, since the likelihood function is nonzero only for a very narrow f-range, applying the scale-invariant Jeffreys' prior would change the posterior only marginally. The prior for the protein structure $\mathbf{R}_j$ of state j was taken from an unbiased MD simulation. Hence, $\pi(\mathbf{R}_j|K)$ is given by a Boltzmann factor of the force field energy V_{ff}, marginalized with respect to all solvent coordinates $\mathbf{r}_{\mathrm{sol}}$ (water and ions), $\pi(\mathbf{R}_i|K) \propto \int \mathrm{d}\mathbf{r}_{\mathrm{sol}} \exp[-\beta V_{\mathrm{ff}}(\mathbf{R}_i, \mathbf{r}_{\mathrm{sol}}, K)]$. Assuming no prior information on the weights, $\pi(\mathbf{w}|K)$ was taken as a flat Dirichlet distribution. For the buffer density mismatch $\delta\rho_{\mathrm{buf}}$, a Gaussian prior was taken as $\pi(\delta\rho_{\mathrm{buf}}) \propto \exp[-\delta\rho_{\mathrm{buf}}^2/(2\epsilon_{\mathrm{buf}}^2)]$. Here, ϵ_{buf} is a free parameter that quantifies the uncertainty of the density of an experimental buffer. Typical values for ϵ_{buf} would be 0.1 to 0.5% of the density of water, yet we found that the choice for ϵ_{buf} had only a small effect on $p(\mathbf{R}, \mathbf{w})$.

On-the-fly calculation of SAXS curves from MD simulations

The buffer-subtracted SAXS curves were derived by explicit-solvent calculations, as described previously [18, 21]. Because the explicit solvent provides an accurate model for the hydration layer and excluded solvent, these calculations did not require any solvent-related fitting parameters, in contrast to implicit-solvent SAXS calculations.

In short, a spatial envelope was constructed around the protein at a distance of at least 8Å from all protein atoms (Fig 1). All protein and solvent atoms within the envelope were taken into account for the calculation of the SAXS curve, as visualized in Fig 1. Likewise, scattering contributions from the excluded solvent were computed from solvent atoms within the envelope taken from a pure-water MD simulation. A memory time constant of $\tau = 500$ ps was applied during both LBP and Hsp90 simulations. The orientational average (or spherical quadrature) was conducted numerically using 1200 $\mathbf{q}$-vectors per absolute value of q, distributed by the spiral method. During SAXS refinement simulations, the SAXS curves were updated on-the-fly every 0.5 ps. The statistical uncertainty σ_{c} of the calculated intensity was computed by applying standard Gaussian error propagation to the SAXS intensity calculations we described previously [21]. After averaging over a few hundred MD frames, σ_{c} is typically small compared to the other uncertainties that contribute to the likelihood function (σ_{buf} and σ_{exp}).

The SAXS curves of the purely open and purely closed states of LBP (Fig 2, solid lines) were computed from 100-nanosecond free simulations of the open and closed state.

Marginalizing out the fitting parameters *f* and *c*

The two fitting parameters f, corresponding to the absolute scale of the SAXS curve, and the offset c, can be marginalized out analytically at the level of the likelihood. Assuming Gaussian errors, we take for the likelihood

$$L(D|\mathbf{R}, \mathbf{w}, f, c, \delta\rho_{\mathrm{buf}}, K) \propto \exp\left[-\frac{1}{2}\zeta\chi^2\right] \tag{5}$$

with

$$\chi^2 = \sum_{i=1}^{N_q} \tau_{i,f}[I_{\mathrm{c}}(q_i, \mathbf{R}, \mathbf{w}) - (fI_{\mathrm{exp}}(q_i) + c)]^2, \tag{6}$$

where we introduced the symbol $\zeta = N_{\mathrm{indep}}/N_q$, as well as the precision of the i^{th} q-point as

follows:

$$\tau_{i,f} = \frac{1}{\sigma^2(q_i)} = \frac{1}{f^2\sigma^2_{\mathrm{exp}}(q_i) + \sigma^2_{\mathrm{c}}(q_i) + \sigma^2_{\mathrm{buf}}(q_i;\delta\rho_{\mathrm{buf}})}. \quad (7)$$

Here, we used that the uncertainties from the experiment σ_{exp}, from the calculation σ_{c}, and from the buffer subtraction σ_{buf} are independent, suggesting that the respective variances add up to the total variance $\sigma^2(q_i)$. The precision $\tau_{i,f}$ depends on the fitted scale f because the experimental errors σ_{exp} must be scaled simultaneously with the experimental intensities I_{exp}. To allow us to marginalize out the scale f analytically, we use that the errors in the small-angle regime are much smaller than the intensities, suggesting that purely values of f close to it's maximum-likelihood estimate f_{ml} contribute to the marginalized likelihood. As a consequence, replacing f by f_{ml} in eq 7 has only a small effect on the marginalized likelihood. We therefore use for the precision in the following

$$\tau_i = [f^2_{\mathrm{ml}}\sigma^2_{\mathrm{exp}}(q_i) + \sigma^2_{\mathrm{calc}}(q_i) + \sigma^2_{\mathrm{buf}}(q_i;\delta\rho_{\mathrm{buf}})]^{-1}. \quad (8)$$

In the first calculation step, while f_{ml} is still unknown, it may be simply estimated from the non-weighted averages of the calculated and experimental intensities, following $f_{\mathrm{ml}} \approx \Sigma_i I_c(q_i)/\Sigma_i I_{\mathrm{exp}}(q_i)$.

To keep the nomenclature clear, let us introduce additional symbols. Let $T := \sum_{i=1}^{N_q} \tau_i$ denote the sum over all precisions. The τ-weighted average over q-points is

$$\langle X \rangle = T^{-1}\sum_{i=1}^{N_q} \tau_i X(q_i). \quad (9)$$

With the last definition, the τ-weighted variances of the calculated and experimental SAXS intensities are

$$s_c^2 = \langle I_c^2 \rangle - \langle I_c \rangle^2 \quad (10)$$

$$s_{\mathrm{exp}}^2 = \langle I_{\mathrm{exp}}^2 \rangle - \langle I_{\mathrm{exp}} \rangle^2, \quad (11)$$

respectively, and the τ-weighted Pearson correlation coefficient between the calculated and experimental data points is

$$P = \frac{\langle I_c\, I_{\mathrm{exp}} \rangle - \langle I_c \rangle \langle I_{\mathrm{exp}} \rangle}{s_c\, s_{\mathrm{exp}}}. \quad (12)$$

The maximum likelihood estimates for the fitting parameters f and c are

$$f_{\mathrm{ml}} = P\frac{s_c}{s_{\mathrm{exp}}} \quad (13)$$

$$c_{\mathrm{ml}} = \langle I_c \rangle - f_{\mathrm{ml}} \langle I_{\mathrm{exp}} \rangle, \quad (14)$$

respectively. The residual between I_c and I_{exp} that cannot be fitted by the parameters f and c is

$$\hat{\chi}^2 = T[s_c^2 - (f_{\mathrm{ml}} s_{\mathrm{exp}})^2] \quad (15)$$

$$= T\langle [I_c - (f_{\mathrm{ml}} I_{\mathrm{exp}} + c_{\mathrm{ml}})]^2 \rangle. \quad (16)$$

The last equality is derived using eqs 10 to 14. The likelihood L_{marg} marginalized with respect to fitting parameters f and c is obtained by integrating over f and c. Since no prior information on f and c is available, we assumed flat prior distributions, $\pi(f) = \pi(c) = 1$. A straightforward calculation yields:

$$\begin{aligned} & L_{\mathrm{marg}}(D|\mathbf{R}, \mathbf{w}, \sigma_{\mathrm{buf}}, K) \\ & \quad \propto \int \mathrm{d}f \int \mathrm{d}c\, L(D|\mathbf{R}, \mathbf{w}, f, c, \sigma_{\mathrm{buf}}, K)\pi(f)\pi(c) \\ & \quad \propto \frac{1}{Ts_{\mathrm{exp}}} \exp\left(-\frac{1}{2}\zeta\hat{\chi}^2\right) \end{aligned} \tag{17}$$

Here, we dropped the normalization factors and other constants of the likelihood because these only lead to an irrelevant constant offset in the experiment-derived energies.

Force calculations

In order to sample the posterior distribution using Newtonian dynamics, L_{marg} is reformulated as its energy analogue

$$E_{\mathrm{exp}} = -\beta^{-1} \ln L_{\mathrm{marg}}. \tag{18}$$

Using eqs 10 through 18, the experiment-derived force on atom ℓ of state j to is

$$\mathbf{F}_{j,\ell} = -\frac{\partial}{\partial \mathbf{r}_{j,\ell}} E_{\mathrm{exp}} \tag{19}$$

$$= -\beta^{-1}\zeta \sum_{i=1}^{N_q} \tau_i \left[I_c(q_i) - (f_{\mathrm{ml}} I_{\mathrm{exp}}(q_i) + c_{\mathrm{ml}}) \right] \frac{\partial I_c(q_i, \mathbf{R}, \mathbf{w})}{\partial \mathbf{r}_{j,\ell}}, \tag{20}$$

where $\mathbf{r}_{j,\ell}$ denotes the Cartesian coordinates of atom ℓ in state j ($j = 1, \ldots, N$). In general, the calculated SAXS intensity I_c is a weighted average over the intensities of the N states:

$$I_c(q_i, \mathbf{R}, \mathbf{w}) = \sum_{j=1}^{N} w_j I(q_i, \mathbf{R}_j), \tag{21}$$

where w_j and $I(q_i, \mathbf{R}_j)$ denote the normalized weight ($\sum_j w_j = 1$) and the SAXS intensity of state j, respectively. Following eq 21, the derivative of I_c with respect to $\mathbf{r}_{j,\ell}$ depends purely on the SAXS intensity of state j:

$$\frac{\partial I_c(q_i, \mathbf{R}, \mathbf{w})}{\partial \mathbf{r}_{j,\ell}} = w_j \frac{\partial I(q_i, \mathbf{R}_j)}{\partial \mathbf{r}_{j,\ell}}. \tag{22}$$

Note that, for the simulations conducted in this study, one closed state ($j = 1$) was assumed to adopt a fixed know structure, whereas a second open state ($j = 2$) was refined against SAXS data. Hence, the forces $\mathbf{F}_{j,\ell}$ were purely evaluated for the second flexible state. However, the equations presented above are suitable for simultaneously refining multiple states against SAXS data. The derivative $\partial I(q_i, \mathbf{R}_j)/\partial \mathbf{r}_{j,\ell}$ was computed as described previously [18, 21].

Marginalizing out the scale *f* only

For the simulations of this study, we applied the likelihood function defined in eqs 5 and 6, using both the absolute scale *f* and the constant offset *c* as unknown fitting parameters. However, there may be applications for which the fitting of a constant offset *c* is undesirable. Hence, for the sake of completeness, we report the expressions for marginalizing out purely the absolute scale *f*.

Then, the likelihood takes the form of eqs 5 and 6 with the parameter *c* set to zero. The maximum-likelihood estimate for the scale *f* evaluates to $f'_{\mathrm{ml}} = \langle I_c\, I_{\mathrm{exp}}\rangle / \langle I^2_{\mathrm{exp}}\rangle$, and the residual between I_c and I_{exp} changes to

$$\hat{\chi}'^2 = T(\langle I_c^2\rangle - \langle I_c I_{\mathrm{exp}}\rangle^2 / \langle I^2_{\mathrm{exp}}\rangle) = T\langle [I_c - f'_{\mathrm{ml}} I_{\mathrm{exp}}]^2\rangle. \tag{23}$$

The marginalized likelihood is

$$L'_{\mathrm{marg}} \propto \frac{1}{[T\langle I_{\mathrm{exp}}\rangle]^{1/2}} \exp\left(-\frac{1}{2}\zeta\hat{\chi}'^2\right), \tag{24}$$

and the force on atom ℓ of state *j*

$$\mathbf{F}_{j,\ell} = -\beta^{-1}\zeta\sum_{i=1}^{N_q}\tau_i\left[I_c(q_i) - f_{\mathrm{ml}} I_{\mathrm{exp}}(q_i)\right]\frac{\partial I_c(q_i)}{\partial \mathbf{r}_{j,\ell}}. \tag{25}$$

Monte-Carlo moves for $\delta\rho_{\mathrm{buf}}$ and weights

The weights of the *N* states ($N = 2$ in this study), as well as the uncertainty of the buffer density $\delta\rho_{\mathrm{buf}}$ were sampled using Gibbs sampling, that is, using Monte-Carlo (MC) moves with all other parameters fixed. At each time step at which the SAXS intensities were updated (0.5 ps in this study), 20 rounds of MC moves of $\delta\rho_{\mathrm{buf}}$ and w_{open} were conducted. In each round, 20 MC moves of $\delta\rho_{\mathrm{buf}}$ were conducted (at fixed w_{open}), followed by 20 MC moves of w_{open} (at fixed $\delta\rho_{\mathrm{buf}}$). Typical posteriors of the parameter $\delta\rho_{\mathrm{buf}}$ are shown in S3 Fig.

Proposed MC moves of $\delta\rho_{\mathrm{buf}}$ were taken from a uniform distribution in the interval $[0, 6\epsilon_{\mathrm{buf}})$. Proposed MC moves for the weight vector $\mathbf{w} = (w_1, \ldots, w_N)$ were taken from a flat Dirichlet distribution. Hence, proposed $\mathbf{w}$ satisfied $\sum_{i=1}^{N} w_i = 1$ and were uniformly distributed over the standard $(N-1)$-simplex, that is, the prior $\pi(\mathbf{w})$ was a constant. Such $\mathbf{w}$ were drawn from the flat Dirichlet distribution by randomly partitioning the interval [0, 1], as follows:

1. Take random numbers R_i ($i = 1, \ldots, N-1$) from a uniform distribution in the interval [0, 1]. In addition, set $R_0 = 0$ and $R_N = 1$.
2. Sort the R_i in increasing order.
3. Take proposed weights as $w_i = R_i - R_{i-1}$.

We noticed that restricting the sampling of w_i to the interval [0, 1] may lead to artifacts at "edge" of the $(N-1)$-simplex, presumably as a consequence of the weighted running averages used for computing SAXS curves on-the-fly during MD simulations [18]. To avoid a boundary in the physically relevant weights space, the sampled weight space was extended to unphysical but mathematically well-defined weights slightly outside the $(N-1)$-simplex (outside the interval [0, 1] in case of $N = 2$). This was achieved by scaling the proposed weight vector $\mathbf{w}$,

followed by a shift along the vector with all elements equal to unity, $\mathbf{j} = (1, \ldots, 1)$, as follows:

$$\mathbf{w}' = (1 + \xi N)\mathbf{w} - \xi\,\mathbf{j}. \tag{26}$$

The parameter ξ was set to 0.1 in this study. This transformation keeps the prior of $\mathbf{w}'$ uniform on the $(N-1)$-simplex, and it keeps the weight vector normalized ($\sum_{i=1}^{N} w_i' = 1$). However, it allows one to draw samples of w_i' from the interval $[-\xi, 1 + N\xi - \xi]$. For $N = 2$, for instance, samples of w_i' are drawn from the interval $[-\xi, 1 + \xi]$.

The proposed MC move was accepted with probability P_{accept} according to the Metropolis algorithm,

$$P_{\text{accept}} = \min\{1, p'_{\text{marg}}/p_{\text{marg}}\}, \tag{27}$$

where the prime indicates the posterior after the MC move. Further, the symbol p_{marg} denotes the posterior distribution after marginalizing out the fitting parameters, which is given by

$$p_{\text{marg}}(\mathbf{R}, \mathbf{w}, \delta\rho_{\text{buf}}|D, K) \propto L_{\text{marg}}(D|\mathbf{R}, \mathbf{w}, \delta\rho_{\text{buf}}, K)\pi(\mathbf{R}|K)\pi(\mathbf{w}|K)\pi(\delta\rho_{\text{buf}}|K). \tag{28}$$

For each MC move p_{marg} was evaluated using eq 17 as well as the priors for $\mathbf{w}$ (a constant in this study) and $\pi(\sigma_{\text{buf}})$ (a Gaussian in this study, see section on prior distributions).

Umbrella sampling along open/closed weights

Obtaining a (reasonably) converged posterior distribution as a function of weights and protein coordinates would require very long simulations. To ensure exhaustive sampling of the weights space and, hence, to accelerate the convergence of the posterior, we used umbrella sampling along the weights [32]. Further, umbrella sampling is technically convenient because it allows the calculation of the posterior from a set of independent simulations.

For the two-state refinement used here, one-dimensional umbrella sampling was sufficient. Accordingly, the weight of the open state w_{open} was decomposed into $N_{\text{win}} = 11$ umbrella windows $w_{\text{open}}^{(k)} = \{0, 0.1, \ldots, 1.0\}$ ($k = 1, \ldots, N_{\text{win}}$). During MC moves of the weights, a harmonic umbrella potential was applied $V_k^{(b)} = f_w(w_{\text{open}} - w_{\text{open}}^{(k)})^2/2$ or, equivalently, the MC moves were accepted or rejected based on the biased posterior

$$p_{\text{marg},k}^{(b)} = p_{\text{marg}}\, \mathrm{e}^{-\beta V_k^{(b)}}. \tag{29}$$

An umbrella force constant of $f_w = 1000$ kJ/mol was applied. A typical set of umbrella histograms is shown in S4 Fig, demonstrating sufficient overlap between neighboring histograms. After the simulations had finished, the umbrella windows were combined to the unbiased posterior using the weighted histogram analysis method (WHAM), as implemented in the g_wham software [62, 63].

Schematic overview of the algorithm

Fig 7 visualizes the algorithm used to compute the posteriors. Accordingly, the simulation system is set up from the initial coordinates $\mathbf{R}$, and initial values for the weights $\mathbf{w}$ and the buffer density mismatch $\delta\rho_{\text{buf}}$ are defined. The system is freely simulated for τ (the memory time constant for on-the-fly SAXS calculations [18]), using purely the MD force field V_{ff}. The free simulation is required is required because, using the explicit-solvent SAXS predictions, the SAXS curve cannot be computed from a single frame but instead requires averaging over solvent fluctuations. Within the free simulation, an initial estimate for the calculated SAXS intensity $I_c(q_i, \mathbf{R}, \mathbf{w})$ is obtained. A typical value for τ is 300 ps, suggesting that the computed SAXS

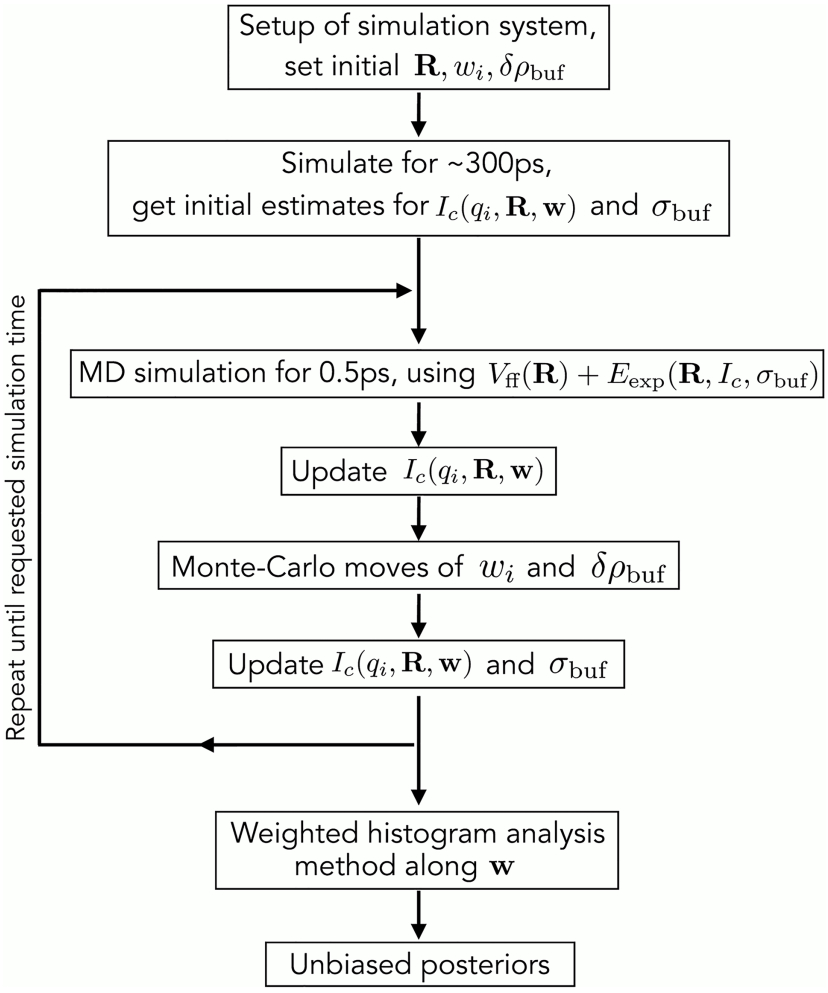

Fig 7. Overview of the algorithm used to compute the posteriors. For more details, see text.

https://doi.org/10.1371/journal.pcbi.1005800.g007

curves account for fluctuations on the several hundred picosecond time scale. Subsequently, the experiment-derived energy E_{exp} is gradually turned on within the following τ.

The following steps are repeated until the requested simulation time is reached for each umbrella window along the weights: (i) MD simulation using forces derived from the hybrid energy, i.e., using the potential $V_{\mathrm{ff}} + E_{\mathrm{exp}}$; (ii) update of $I_c(q_i, \mathbf{R}, \mathbf{w})$ based on the current MD frame and using a cumulative weighted average [18], as previously suggested for NMR refinement [64]; (iii) a few hundred MC moves of weights w_i and $\delta\rho_{\mathrm{buf}}$ (see above); (iv) update $I_c(q_i, \mathbf{R}, \mathbf{w})$ with the final weights, and update the systematic error σ_{buf} with the final $\delta\rho_{\mathrm{buf}}$, as described previously [18]. After the simulations from all umbrella windows have finished, the biased posteriors from all windows are combined into the unbiased posterior using WHAM [62].

Preparation of structures for MD simulations

The crystal structures of the apo and holo states of LBP were taken from the protein data bank (PDB codes 1USG and 1USI [35]). For the simulation of Hsp90 the structure of ATP-bound yeast Hsp90 was used (PDB code 2CG9) [36]. The co-chaperone proteins SBA1 and ATP

ligands were removed from the structure of HspP90 and leucine ligand was removed from the LBP structure. Flexible linkers missing in the Hsp90 crystal structure were added using Modeller [65]. A swap of the N-terminal β-strand (residues 1-9), which prevented the opening of the protein, was removed using the Coot software [66].

The structures were placed in simulation boxes of a rhombic dodecahedron with distance between the protein and the box surface of 1.5 and 4 nm for LBP and Hsp90, respectively. The systems were solvated by explicit water. Sodium and chloride ions were added to obtain a salt concentration of 100 mM. Here, the number of sodium and chloride ions was adjusted to neutralize the system. The energies of the systems were minimized using the steepest descent algorithm for 2000 steps. Subsequently, the systems were equilibrated with position restraints on the backbone atoms for 10 and 20 ns for LBP and Hsp90, respectively.

To obtain an initial open structure of Hsp90, we carried out pulling simulations along the distance of the two N-terminal domains. Accordingly, the center-of-mass distance between the two N-terminal domains was increased from 4 nm to 8 nm within 40 ns, using a pulling speed of 0.1 nm/ns. The obtained open structure was resolvated in a larger simulation box with the distance between the protein and the box surface of 3 nm, and the structure was equilibrated for another 20 ns with position restraints on the backbone atoms. The final structure was used as a starting structure for SAXS refinement. The Hsp90 system contained approximately 1.5×10^6 atoms.

MD parameters

Standard MD simulations were performed using the GROMACS simulation software (version 4.6) [67]. All SAXS calculations were done with an in-house modification of GROMACS 4.6, which is available from the authors upon request. Protein interactions of LBP and Hsp90 were taken from the CHARMM27 and CHARMM22* forcefields, respectively [68, 69], and water was modeled by the TIP3P potential [70]. Hydrogen atoms of the proteins were modeled as virtual interaction sites allowing an integration timestep of 4 fs. Electrostatic interactions were treated with the particle-mesh Ewald scheme [71, 72]. The cutoff of 1.2 nm was applied to the direct-space Coulomb and Lennard-Jones interactions. The bond lengths and angles of water molecules were constrained with the SETTLE algorithm [73], and all other bonds were constrained with LINCS [74]. The pressure was set to 1 bar using the Berendsen barostat (τ = 1 ps) [75]. During equilibration runs, the temperature was controlled at 300 K with the Berendsen thermostat (τ = 0.5 ps) [75]. During SAXS-driven simulations, in contrast, a tight stochastic dynamics integration scheme was applied, motivated from the fact that SAXS-driven MD is not strictly energy conservative [76].

SAXS refinement simulations

For LBP simulations, the target curves for the refinement were modeled from calculated SAXS curves of the closed state $I_{\text{closed}}(q)$ (Fig 2C, solid dark blue curve) and open state $I_{\text{open}}(q)$ (Fig 2C, solid light blue curve), as follows:

$$I_{\text{exp},w}(q) = w_{\text{open}}^{\text{exp}} I_{\text{open}}(q) + (1 - w_{\text{open}}^{\text{exp}}) I_{\text{closed}}(q) \tag{30}$$

In this study, we tested ensemble refinement against SAXS data computed with the following $w_{\text{open}}^{\text{exp}}$: 0, 25%, 50%, 75%, or 100% (Fig 2C, solid and dashed curves). Hence, since $I_{\text{exp},w}(q)$ was computed theoretically, the true weight of the open state was known, allowing us

1. to validate that the SAXS-guided refinement starting from the closed state is capable of reproducing the true weight ($w_{\text{open}}^{\text{exp}}$) and the true structure of the open state (used to compute $I_{\text{open}}(q)$), and
2. to derive the uncertainty (or ambiguity) of the weight and structure in the light of the SAXS curve and the MD force field, as given by the width of the posterior distributions.

All simulations of LBP were started from the closed state (Fig 2A). The simulations were coupled to the target SAXS curve at $N_q = 25$ q-points, which were evenly distributed between 0 and 8 nm^{-1}. The two-state ensemble refinement was conducted using umbrella sampling along the weight w_{open} of the open state (see above). Each umbrella window was simulated for 40 ns, where the first 2 ns were removed for equilibration. The posterior distributions of w_{open} and of the interdomain distance derived from these simulations are presented in Figs 2D, 3A and S1A Fig.

For comparison, a single state (instead of the ensemble of two states) was refined against each of the five curves $I_{\text{exp},w}(q)$, using five simulations of 10 ns each and removing the first 2 ns for equilibration. S1B Fig presents the posteriors of the interdomain distance d_{NC} resulting from refining a single structure against SAXS curves that, in truth, represent heterogenous open/closed ensemble. Notably, the single-state refinements try to explain those SAXS curves with intermediate (partially open) structures.

For the refinement simulations of Hsp90, the simulations were coupled to the target SAXS curve at $N_q = 30$ q-points, which were evenly distributed between 0.1 and 3 nm^{-1}. The q-range below 0.1 nm^{-1} was omitted because the experimental data exhibited some deviation from the ideal Guinier behaviour. For some umbrella windows, Hsp90 was required to carry out large-scale conformational transitions. To accelerate those transitions, each window was first simulated for 8 ns with a ten-fold increased experiment-derived energy E_{exp}. Subsequently, the simulation of each umbrella window was continued for another 20 ns using the statistically founded E_{exp} that leads to the correct posterior (eq 18). From those simulations, the first 2 ns were removed for further equilibration, and the remaining simulation time was used to compute the posterior. An example of the umbrella histograms along the weight coordinate is shown in S4 Fig. To further improve the sampling close to the maxima of the posteriors, the simulations of the umbrella window at the peak of $p(w_{\text{open}}|D, K)$ plus two neighboring windows were prolonged for another 15 ns.

Supporting information

S1 Fig. Marginalized posteriors of the interdomain distance d_{NC} of the refined open state, taken from two-state and single-state refinement simulations of LBP. (A) Marginalized posteriors of the interdomain distance d_{NC} of the refined open state, taken from two-state refinement simulations of LBP. In ensembles refined against SAXS curves of non-zero open-state content (25:75 through 100:0), the posteriors peak near the physically correct d_{NC} of $\sim$3.25 nm of the open state. In the ensemble refined against the SAXS curve of purely the closed state (0:100), the refined weight of the open state is near zero (Fig 3A of main text), suggesting that the simulation of the open state is hardly restrained by the SAXS curve or, equivalently, is essentially a free simulation. Consequently, the posterior of d_{NC} (A, dark green) is wide and reflects both closed and open states.
(PDF)

S2 Fig. Target SAXS curves (black lines) and calculated SAXS curves of the refined structures and ensembles (red dots) of leucine binding protein. (A/C) Two-state ensemble refinement. (B/D) Single-state refinement. SAXS curves with open weight $\geq$25% were offset for

clarity. The lower row (C/D) shows a close-up view on the small-angle regime.
(PDF)

S3 Fig. Posterior of $\delta\rho_{buf}/\epsilon_{buf}$ during two state refinement of LBP and Hsp90 refinement. For LBP, the ensembles were refined against theoretically computed SAXS curves, thus exhibiting no buffer density mismatch, rationalizing why the posteriors peak near $\delta\rho_{buf} = 0$. For Hsp90, in contrast, the ensembles were refined against experimental data that presumably exhibit some systematic errors, for instance due to a small buffer density mismatch. Hence, the posteriors peak at nonzero $\delta\rho_{buf}$.
(PDF)

S4 Fig. Example of umbrella histograms. Umbrella histograms along the weight of the open state, here taken from the two-state refinement of LBP against the open/closed 50:50 SAXS curve.
(PDF)

S5 Fig. Root mean-square deviation (RMSD) analysis of LBP. Root mean-square deviation (RMSDs) between the C_α atoms of (i) the refined open structure of LBP and (ii) the open LBP structure (similar to Fig 2B). For the RMSD calculations, the open structure was taken as the average structure of the ensembles used to compute the SAXS curve of the open state. Trajectories were taken from umbrella windows of w_{open} close to the peak of the posterior $p(w_{open}|D, K)$. The color indicates the open:closed weights used to compute the target curve (see legend, five target curves in Fig 2C). The RMSD curves demonstrate that, starting from the closed state, LBP rapidly opens and approaches the open structure. Some RMSD fluctuations after longer time (green curves) reflect a smaller twist motion between the N- and C-terminal domains within the open state. Such fluctuations along the twist are expected since, as shown previously [18], the SAXS data restrains the degree of openness but not the twist.
(PDF)

S6 Fig. Analysis of the convergence of posterior distributions with increasing invested simulation time. (A/B) Marginalized posteriors for LBP refined against the SAXS curve with 50:50 open/closed weight (Fig 2D, middle panel), computed from time bins as indicated in the legend. (A) Posterior of the interdomain distance d_{NC} and (B) of the weight of the open state w_{open}. (C) Marginalized posteriors of w_{open} for different liganded states of Hsp90 as indicated in the legend. The thin lines indicate posteriors computed from an increasing number of histograms: 10 equally spaced histograms, the same 10 plus additional three histograms near the posterior maximum, 10 plus 6 additional histograms, and 10 plus 9 additional histograms near the posterior maximum. The similarity between the posteriors suggest that the posteriors are reasonably converged.
(PDF)

S1 Table. Maxima and confidence intervals of w_{open}, taken from $p(w_{open}|D, K)$ of the two-state ensemble refinement of LBP. All numbers in %. The respective posteriors are shown in Fig 3A.
(PDF)

S2 Table. Maxima and confidence intervals of d_{NC}, taken from $p(d_{NC}|D, K)$ of the two-state ensemble refinement of LBP. During two-state refinement, in simulations with non-zero w_{open}, the posterior of the interdomain distance $p(d_{NC}|D, K)$ of the open state peaks at the physically correct $d_{NC} \approx 3.25$ nm. The respective posteriors are shown in S1A Fig. All distances are in nanometers.
(PDF)

S3 Table. Maxima and confidence intervals of d_{NC} taken from $p(d_{NC}|D, K)$ of the single-state refinement of LBP. Refining a single state against SAXS curves that, in truth, represent a heterogenous ensemble of open/closed states, yields posterior distributions that peak at the "mean" interdomain distance $\langle d_{\mathrm{NC}} \rangle = w_{\mathrm{open}} d_{\mathrm{NC}}^{\mathrm{open}} + (1 - w_{\mathrm{open}}) d_{\mathrm{NC}}^{\mathrm{closed}}$, where $d_{\mathrm{NC}}^{\mathrm{open}}$ and $d_{\mathrm{NC}}^{\mathrm{closed}}$ denote the mean interdomain distances of the open and closed states, in free simulations, respectively. The respective posteriors are shown in Fig 1B. All distances are in nanometers.
(PDF)

S4 Table. Maxima and confidence intervals of w_{open} taken from $p(w_{open}|D, K)$ of the two-state ensemble refinement of Hsp90. All numbers in %. The respective posteriors are shown in Fig 4D of the main text.
(PDF)

S1 Archive. Source code of modified version of gromacs which was used for SAXS-driven MD simulations.
(BZ2)

S2 Archive. Setup files for the LBP two-state ensemble refinement.
(BZ)

S3 Archive. Setup files for the HSP90 two-state ensemble refinement.
(BZ)

Acknowledgments

We thank Michael Habeck for many helpful discussions, for support with marginalizing out the fitting parameters, and for critically reading the manuscript. We thank Tobias Madl for sharing the experimental SAXS data of Hsp90, and Thomas Monecke for help with the Coot modeling. We thank Kalina Atkovska for proofreading the manuscript.

Author Contributions

Conceptualization: Roman Shevchuk, Jochen S. Hub.

Formal analysis: Jochen S. Hub.

Funding acquisition: Jochen S. Hub.

Investigation: Roman Shevchuk, Jochen S. Hub.

Methodology: Roman Shevchuk, Jochen S. Hub.

Project administration: Jochen S. Hub.

Software: Roman Shevchuk, Jochen S. Hub.

Supervision: Jochen S. Hub.

Visualization: Roman Shevchuk.

Writing – original draft: Roman Shevchuk, Jochen S. Hub.

Writing – review & editing: Jochen S. Hub.

References

1. Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol. 2009; 5(11):789–796. https://doi.org/10.1038/nchembio.232 PMID: 19841628

2. Schneidman-Duhovny D, Pellarin R, Sali A. Uncertainty in integrative structural modeling. Curr Opin Struct Biol. 2014; 28:96–104. https://doi.org/10.1016/j.sbi.2014.08.001 PMID: 25173450
3. Schröder GF. Hybrid methods for macromolecular structure determination: experiment with expectations. Curr Opin Struct Biol. 2015; 31:20–27. https://doi.org/10.1016/j.sbi.2015.02.016 PMID: 25795086
4. Jaynes ET. Probability theory: The logic of science. Cambridge University Press; 2003.
5. Habeck M, Rieping W, Nilges M. Weighting of experimental evidence in macromolecular structure determination. Proc Nat Acad Sci USA. 2006; 103(6):1756–1761. https://doi.org/10.1073/pnas.0506412103 PMID: 16446450
6. Rieping W, Habeck M, Nilges M. Inferential structure determination. Science. 2005; 309(5732):303–306. https://doi.org/10.1126/science.1110428 PMID: 16002620
7. Scheres SHW. A Bayesian view on cryo-EM structure determination. J Mol Biol. 2012; 415(2):406–418. https://doi.org/10.1016/j.jmb.2011.11.010 PMID: 22100448
8. Cossio P, Hummer G. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. J Struct Biol. 2013; 184(3):427–437. https://doi.org/10.1016/j.jsb.2013.10.006 PMID: 24161733
9. MacCallum JL, Perez A, Dill KA. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. Proc Natl Acad Sci USA. 2015; 112(22):6985–6990. https://doi.org/10.1073/pnas.1506788112 PMID: 26038552
10. Walczak M, Grubmüller H. Bayesian orientation estimate and structure information from sparse single-molecule x-ray diffraction images. Phys Rev E. 2014; 90(2):022714. https://doi.org/10.1103/PhysRevE.90.022714
11. Bonomi M, Camilloni C, Cavalli A, Vendruscolo M. Metainference: A Bayesian inference method for heterogeneous systems. Sci Adv. 2016; 2(1):e1501177. https://doi.org/10.1126/sciadv.1501177 PMID: 26844300
12. Koch MH, Vachette P, Svergun DI. Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. Quart Rev Biophys. 2003; 36(02):147–227. https://doi.org/10.1017/S0033583503003871
13. Putnam CD, Hammel M, Hura GL, Tainer JA. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Quart Rev Biophys. 2007; 40(3):191–285. https://doi.org/10.1017/S0033583507004635
14. Petoukhov MV, Svergun DI. Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophys J. 2005; 89(2):1237–1250. https://doi.org/10.1529/biophysj.105.064154 PMID: 15923225
15. Gorba C, Miyashita O, Tama F. Normal-mode flexible fitting of high-resolution structure of biological molecules toward one-dimensional low-resolution data. Biophys J. 2008; 94(5):1589–1599. https://doi.org/10.1529/biophysj.107.122218 PMID: 17993489
16. Pelikan M, Hura GL, Hammel M. Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys. 2009; 28(2):174. https://doi.org/10.4149/gpb_2009_02_174 PMID: 19592714
17. Zheng W, Tekpinar M. Accurate flexible fitting of high-resolution protein structures to small-angle x-ray scattering data using a coarse-grained model with implicit hydration shell. Biophys J. 2011; 101(12):2981–2991. https://doi.org/10.1016/j.bpj.2011.11.003 PMID: 22208197
18. Chen P, Hub JS. Interpretation of solution X-ray scattering by explicit-solvent molecular dynamics. Biophys J. 2015; 108:2573–2584. https://doi.org/10.1016/j.bpj.2015.03.062 PMID: 25992735
19. Kimanius D, Pettersson I, Schluckebier G, Lindahl E, Andersson M. SAXS-guided metadynamics. J Chem Theory Comput. 2015; 11(7):3491–3498. https://doi.org/10.1021/acs.jctc.5b00299 PMID: 26575782
20. Alexander Björling A, Niebling S, Marcellini M, van der Spoel D, Westenhoff S. Deciphering solution scattering data with experimentally guided molecular dynamics simulations. J Chem Theory Comput. 2015; 11:780–787. https://doi.org/10.1021/ct5009735
21. Chen P, Hub JS. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. Biophys J. 2014; 107:435–447. https://doi.org/10.1016/j.bpj.2014.06.006 PMID: 25028885
22. Chen Pc, Hub JS. Structural Properties of Protein-Detergent Complexes from SAXS and MD Simulations. J Phys Chem Lett. 2015; 6:5116–5121. https://doi.org/10.1021/acs.jpclett.5b02399 PMID: 26637017

23. Ravera E, Sgheri L, Parigi G, Luchinat C. A critical assessment of methods to recover information from averaged data. Phys Chem Chem Phys. 2016; 18(8):5686–5701. https://doi.org/10.1039/C5CP04077A PMID: 26565805

24. Yang S, Blachowicz L, Makowski L, Roux B. Multidomain assembled states of Hck tyrosine kinase in solution. Proc Natl Acad Sci USA. 2010; 107(36):15757–15762. https://doi.org/10.1073/pnas.1004569107 PMID: 20798061

25. Rózycki B, Kim YC, Hummer G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. Structure. 2011; 19(1):109–116. https://doi.org/10.1016/j.str.2010.10.006 PMID: 21220121

26. Bernadó P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. Structural characterization of flexible proteins using small-angle X-ray scattering. J Am Chem Soc. 2007; 129(17):5656–5664. https://doi.org/10.1021/ja069124n PMID: 17411046

27. Hummer G, Köfinger J. Bayesian ensemble refinement by replica simulations and reweighting. J Chem Phys. 2015; 143(24):12B634_1. https://doi.org/10.1063/1.4937786

28. Antonov LD, Olsson S, Boomsma W, Hamelryck T. Bayesian inference of protein ensembles from SAXS data. Phys Chem Chem Phys. 2016; 18(8):5832–5838. https://doi.org/10.1039/C5CP04886A PMID: 26548662

29. Svergun D, Barberato C, Koch MHJ. CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J Appl Crystallogr. 1995; 28(6):768–773. https://doi.org/10.1107/S0021889895007047

30. Grishaev A, Guo L, Irving T, Bax A. Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. J Am Chem Soc. 2010; 132:15484–15486. https://doi.org/10.1021/ja106173n PMID: 20958032

31. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophys J. 2013; 105(4):962–974. https://doi.org/10.1016/j.bpj.2013.07.020 PMID: 23972848

32. Torrie GM, Valleau JP. Monte Carlo free energy estimates using non-Boltzmann sampling: Application to the sub-critical Lennard-Jones fluid. Chem Phys Lett. 1974; 28:578–581. https://doi.org/10.1016/0009-2614(74)80109-0

33. Dwyer MA, Hellinga HW. Periplasmic binding proteins: a versatile superfamily for protein engineering. Curr Opin Struct Biol. 2004; 14(4):495–504. https://doi.org/10.1016/j.sbi.2004.07.004 PMID: 15313245

34. Olah GA, Trakhanov S, Trewhella J, Quiocho FA. Leucine/isoleucine/valine-binding protein contracts upon binding of ligand. J Biol Chem. 1993; 268(22):16241–16247. PMID: 8344909

35. Magnusson U, Salopek-Sondi B, Luck LA, Mowbray SL. X-Ray Structures of the Leucine-Binding Protein Illustrate Conformational Changes and the Basis of Ligand Specificity. J Biol Chem. 2004; 279:8747. https://doi.org/10.1074/jbc.M311890200 PMID: 14672931

36. Ali MMU, Roe SM, Vaughan CK, Meyer P, Panaretou B, Piper PW, et al. Crystal structure of an Hsp90–nucleotide–p23/Sba1 closed chaperone complex. Nature. 2006; 440(7087):1013–1017. https://doi.org/10.1038/nature04716 PMID: 16625188

37. Lorenz OR, Freiburger L, Rutz DA, Krause M, Zierer BK, Alvira S, et al. Modulation of the Hsp90 chaperone cycle by a stringent client protein. Mol Cell. 2014; 53(6):941–953. https://doi.org/10.1016/j.molcel.2014.02.003 PMID: 24613341

38. Shiau AK, Harris SF, Southworth DR, Agard DA. Structural analysis of E. coli hsp90 reveals dramatic nucleotide-dependent conformational rearrangements. Cell. 2006; 127(2):329–340. https://doi.org/10.1016/j.cell.2006.09.027 PMID: 17055434

39. Zhao R, Davey M, Hsu YC, Kaplanek P, Tong A, Parsons AB, et al. Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone. Cell. 2005; 120(5):715–727. https://doi.org/10.1016/j.cell.2004.12.024 PMID: 15766533

40. Pearl LH, Prodromou C, Workman P. The Hsp90 molecular chaperone: an open and shut case for treatment. Biochem J. 2008; 410(3):439–453. https://doi.org/10.1042/BJ20071640 PMID: 18290764

41. Wandinger SK, Richter K, Buchner J. The Hsp90 chaperone machinery. J Biol Chem. 2008; 283(27):18473–18477. https://doi.org/10.1074/jbc.R800007200 PMID: 18442971

42. Taipale M, Jarosz DF, Lindquist S. HSP90 at the hub of protein homeostasis: emerging mechanistic insights. Nature Rev Mol Cell Biol. 2010; 11(7):515–528. https://doi.org/10.1038/nrm2918

43. Lai B, Chin N, Stanek A, Keh W, Lanks K. Quantitation and intracellular localization of the 85K heat shock protein by using monoclonal and polyclonal antibodies. Mol Cell Biol. 1984; 4(12):2802–2810. https://doi.org/10.1128/MCB.4.12.2802 PMID: 6396506

44. Whitesell L, Lindquist SL. HSP90 and the chaperoning of cancer. Nat Rev Cancer. 2005; 5(10):761–772. https://doi.org/10.1038/nrc1716 PMID: 16175177

45. Neckers L. Heat shock protein 90: the cancer chaperone. In: Heat Shock Proteins in Cancer. Springer; 2007. p. 231–252.

46. Street TO, Lavery LA, Agard DA. Substrate binding drives large-scale conformational changes in the Hsp90 molecular chaperone. Mol Cell. 2011; 42(1):96–105. https://doi.org/10.1016/j.molcel.2011.01.029 PMID: 21474071

47. Southworth DR, Agard DA. Species-dependent ensembles of conserved conformational states define the Hsp90 chaperone ATPase cycle. Mol Cell. 2008; 32(5):631–640. https://doi.org/10.1016/j.molcel.2008.10.024 PMID: 19061638

48. Krukenberg KA, Förster F, Rice LM, Sali A, A DA. Multiple conformations of E. coli Hsp90 in solution: insights into the conformational dynamics of Hsp90. Structure. 2008; 16(5):755–765. https://doi.org/10.1016/j.str.2008.01.021 PMID: 18462680

49. Zierer BK, Rübbelke M, Tippel F, Madl T, Schopf FH, Rutz DA, et al. Importance of cycle timing for the function of the molecular chaperone Hsp90. Nat Struct Mol Biol. 2016; 23(11):1020–1028. https://doi.org/10.1038/nsmb.3305 PMID: 27723736

50. Hellenkamp B, Wortmann P, Kandzia F, Zacharias M, Hugel T. Multi-domain structure and correlated dynamics determined by self-consistent FRET networks. Nat Methods. 2017; 14(2):174. https://doi.org/10.1038/nmeth.4081 PMID: 27918541

51. Krukenberg KA, Böttcher UM, Southworth DR, Agard DA. Grp94, the endoplasmic reticulum Hsp90, has a similar solution conformation to cytosolic Hsp90 in the absence of nucleotide. Prot Sci. 2009; 18(9):1815–1827. https://doi.org/10.1002/pro.191

52. Franke D, Svergun DI. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. J Appl Cryst. 2009; 42(2):342–346. https://doi.org/10.1107/S0021889809000338

53. Hessling M, Richter K, Buchner J. Dissection of the ATP-induced conformational cycle of the molecular chaperone Hsp90. Nat Struct Mol Biol. 2009; 16(3):287–293. https://doi.org/10.1038/nsmb.1565 PMID: 19234467

54. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic validation of protein force fields against experimental data. PLoS One. 2012; 7(2):e32131. https://doi.org/10.1371/journal.pone.0032131 PMID: 22384157

55. Rauscher S, Gapsys V, Gajda MJ, Zweckstetter M, de Groot BL, Grubmüller H. Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. J Chem Theor Comp. 2015; 11(11):5513–5524. https://doi.org/10.1021/acs.jctc.5b00736

56. Knight CJ, Hub JS. WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. Nucleic Acids Res. 2015; 43:W225–W230. https://doi.org/10.1093/nar/gkv309 PMID: 25855813

57. Fisher CK, Huang A, Stultz CM. Modeling intrinsically disordered proteins with Bayesian statistics. J Am Chem Soc. 2010; 132(42):14919–14927. https://doi.org/10.1021/ja105832g PMID: 20925316

58. Borgia A, Zheng W, Buholzer K, Borgia MB, Schüler A, Hofmann H, et al. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. J Am Chem Soc. 2016; 138(36):11714–11726. https://doi.org/10.1021/jacs.6b05917 PMID: 27583570

59. Pitera JW, Chodera JD. On the use of experimental observations to bias simulated ensembles. J Chem Theor Comp. 2012; 8(10):3445–3451. https://doi.org/10.1021/ct300112v

60. Roux B, Weare J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. J Chem Phys. 2013; 138(8):084107. https://doi.org/10.1063/1.4792208 PMID: 23464140

61. White AD, Voth GA. Efficient and minimal method to bias molecular simulations with experimental data. J Chem Theory Comput. 2014; 10(8):3023–3030. https://doi.org/10.1021/ct500320c PMID: 26588273

62. Kumar S, Bouzida D, Swendsen RH, Kollman PA, Rosenberg JM. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comp Chem. 1992; 13:1011–1021. https://doi.org/10.1002/jcc.540130812

63. Hub JS, de Groot BL, van der Spoel D. g_wham–A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. J Chem Theory Comput. 2010; 6:3713–3720. https://doi.org/10.1021/ct100494z

64. Torda AE, Scheek RM, Van Gunsteren WF. Time-dependent distance restraints in molecular dynamics simulations. Chem Phys Lett. 1989; 157(4):289–294. https://doi.org/10.1016/0009-2614(89)87249-5

65. Webb B, Sali A. Comparative protein structure modeling using Modeller. Curr Protoc Bioinform. 2014; p. 5–6.

66. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D. 2010; 66(4):486–501. https://doi.org/10.1107/S0907444910007493 PMID: 20383002

67. Pronk S, Páll S, Schulz R, Larsson P, Bjelkmar P, Apostolov R, et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. Bioinformatics. 2013; 29(7):845–854. https://doi.org/10.1093/bioinformatics/btt055 PMID: 23407358

68. Piana S, Lindorff-Larsen K, Shaw DE. How robust are protein folding simulations with respect to force field parameterization? Biophys J. 2011; 100(9):L47–L49. https://doi.org/10.1016/j.bpj.2011.03.051 PMID: 21539772

69. MacKerell AD, Feig M, Brooks CL. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem. 2004; 25(11):1400–1415. https://doi.org/10.1002/jcc.20065 PMID: 15185334

70. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983; 79:926–935. https://doi.org/10.1063/1.445869

71. Darden T, York D, Pedersen L. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. J Chem Phys. 1993; 98:10089–10092. https://doi.org/10.1063/1.464397

72. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh ewald potential. J Chem Phys. 1995; 103:8577–8592. https://doi.org/10.1063/1.470117

73. Miyamoto S, Kollman PA. SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithms for Rigid Water Models. J Comp Chem. 1992; 13:952–962. https://doi.org/10.1002/jcc.540130805

74. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A Linear Constraint Solver for Molecular Simulations. J Comp Chem. 1997; 18:1463–1472. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12%3C1463::AID-JCC4%3E3.0.CO;2-H

75. Berendsen HJC, Postma JPM, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys. 1984; 81:3684–3690. https://doi.org/10.1063/1.448118

76. van Gunsteren WF, Berendsen HJC. A Leap-Frog Algorithm for Stochastic Dynamics. Mol Sim. 1988; 1:173–185. https://doi.org/10.1080/08927028808080941