JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2019, 15, 5103−5115

pubs.acs.org/JCTC

SAXS-Restrained Ensemble Simulations of Intrinsically Disordered Proteins with Commitment to the Principle of Maximum Entropy

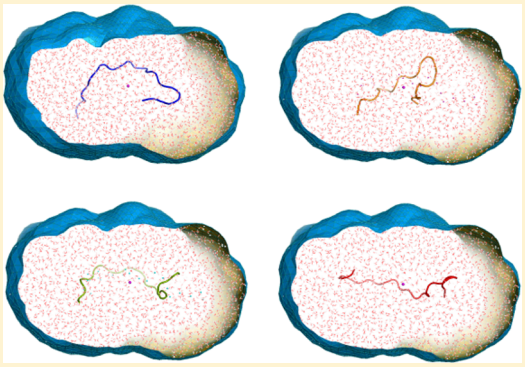
Markus R. Hermann[†] and Jochen S. Hub*[,‡]

[†]Institute for Microbiology and Genetics, Georg-August-Universität Göttingen, 37077 Göttingen, Germany

[‡]Theoretical Physics and Center for Biophysics, Saarland University, Campus E2 6, 66123 Saarbrücken, Germany

Supporting Information

ABSTRACT: Intrinsically disordered proteins (IDPs) play key roles in biology and disease, rationalizing the wide interest in deriving accurate solution ensembles of IDPs. Molecular dynamics (MD) simulations of IDPs often suffer from force-field inaccuracies, suggesting that simulations must be complemented by experimental data to obtain physically correct ensembles. We present a method for integrating small-angle X-ray scattering (SAXS) data on-the-fly into MD simulations of disordered systems, with the aim to bias the conformational sampling toward agreement with ensemble-averaged SAXS data. By coupling a set of parallel replicas to the data and following the principle of maximum entropy, this method applies only a minimal bias. Using the RS peptide as a test case, we analyze the influence of (i) the number of parallel replicas, (ii) the scaling of the force constant for the SAXS-derived biasing energy with the number of parallel replicas, and (iii) the force field. The refined ensembles are cross-validated against experimental $^3J_{\mathrm{HN\text{-}H}\alpha}$ couplings and further compared in terms of C_α distance maps and secondary structure content. Remarkably, we find that the applied force field only has a small influence on the SAXS-refined ensemble, suggesting that incorporating SAXS data into MD simulations may greatly reduce the force-field bias.



INTRODUCTION

Many proteins do not adopt a single, well-defined structure in solution, but instead adopt heterogeneous ensembles. These ensembles are often modulated by the environment, for instance, by varying ion concentrations, temperatures, or via interactions with other molecules. This sensitivity to external stimuli has important implications on their biological function. As such, it is widely accepted that understanding the function of proteins on the molecular level requires understanding of their conformational ensembles.[1] Heterogeneous ensembles are particularly relevant for understanding intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs), which represent roughly 30% of the eukaryotic proteome[2−6] and which play key roles in pathological conditions such as amyloidoses, neurodegenerative diseases, and cancer.[7] However, structural experimental data provide a highly reduced view of protein ensembles. X-ray crystallography or cryo-electron microscopy typically yield only a single yet highly detailed structure, possibly complemented by information on local fluctuations modeled by the Debye−Waller factor. In contrast, nuclear magnetic resonance (NMR) and small-angle X-ray scattering (SAXS) probe the overall ensemble;[8,9] however, NMR and SAXS data represent ensemble averages, and thus likewise provide only a reduced view on the underlying ensemble.

Complementary to experiments, molecular dynamics (MD) simulations generate ensembles of proteins and peptides. The accuracy of MD simulations is often tested by back-calculating ensemble-averaged experimental data from the MD trajectories. If the back-calculated data agree with experiment, there is reason to believe that the simulation generated a reasonably correct ensemble. However, free, unbiased MD simulations frequently disagree with experimental data owing to sampling problems or force-field limitations. For instance, the simulations of IDPs were found to be strongly force-field-dependent.[10−18] Hence, MD simulations have been combined with ensemble-averaged experimental data with the aim to overcome sampling and force-field limitations, and thus, to obtain physically correct conformational ensembles, as discussed in several excellent reviews.[19−26] The data may be used to improve the simulation ensemble either by (i) reweighting the ensemble a posteriori[27−38] or by (ii) restraining the simulation on-the-fly to the experimental data.[19,39−56]

Ensemble reweighting requires that the unbiased ensemble (before reweighting) contains all relevant states. Consequently, the method requires exhaustive sampling. If the sampling is incomplete, i.e., if the unbiased simulation visits only part of the relevant conformational space, the reweighted ensemble may be dominated by only a few structures with high weights.[57] This limitation explains why reweighting has been

Received: April 6, 2019
Published: August 11, 2019

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

often used with simplified, computationally efficient physical models such as coarse-grained models.[33] Similar problems may occur if relevant states were rarely visited due to an artifact in the force field—for instance, if the force field favors overly collapsed conformations of an IDP.[10,58]

Alternatively, an experiment-derived biasing energy $E_{exp}(\mathbf{R}; I_{exp})$ may be used to restrain the simulation on-the-fly into conformations $\mathbf{R}$ that agree with ensemble-averaged data I_{exp}. Here, the simplest approach would be to restrain a single simulation replica with a harmonic restraint to the data. For SAXS data, where experimental scattering intensities $I_{exp}(q_i)$ are available for different momentum transfers q_i, a harmonic restraint may be expressed as

$$E^{(1)}_{exp}(\mathbf{R}; I_{exp}) = \frac{k_r k_B T}{n_q} \sum_{i=1}^{n_q} \frac{[I_c(q_i, \mathbf{R}) - I_{exp}(q_i)]^2}{\sigma_i^2} \quad (1)$$

where k_r is a (unitless) force constant, k_B and T denote the Boltzmann constant and temperature, respectively, and $I_c(q_i, \mathbf{R})$ is the scattering intensity back-calculated from simulation frames $\mathbf{R}$. σ_i denotes the overall uncertainty of ith scattering intensity, which should include uncertainties from both the experiment and from the forward model $I_c(q_i, \mathbf{R})$. The division by n_q ensures that the energy is approximately invariant to rebinning. Together with the energy from the force field $E_{MD}(\mathbf{R})$, this results in a hybrid energy

$$E_{hybrid}(\mathbf{R}; I_{exp}) = E_{MD}(\mathbf{R}) + E^{(1)}_{exp}(\mathbf{R}; I_{exp}) \quad (2)$$

However, this protocol is only appropriate if the ensemble may be approximated by a single, most representative structure.[59−61] In the case of heterogeneous ensembles as adopted by IDPs and IDRs, it is not meaningful to compare the scattering intensity of a single structure $I_c(q_i, \mathbf{R})$ with the ensemble-averaged experimental data $I_{exp}(q_i)$, as the latter represents a diverse range of structures. Instead, because the numerator in eq 1 describes a like-for-like comparison, the ensemble-averaged experimental data I_{exp} should be subtracted against a scattering intensity $\int p(\mathbf{R}) I_c(q_i; \mathbf{R}) d\mathbf{R}$ that is averaged over the simulated ensemble distribution $p(\mathbf{R})$.

A statistically founded procedure for updating simulated ensembles with experimental data is provided by Jaynes' maximum entropy principle.[33,55,62−66] The principle requests that an unbiased ensemble distribution $p_0(\mathbf{R})$ should be modified as minimally as possible into a biased distribution $p_1(\mathbf{R})$ that explains the data, while any bias that is not supported by the data should be strictly avoided. Mathematically, the updated ensemble $p_1(\mathbf{R})$ and the original, unbiased ensemble $p_0(\mathbf{R})$ should be as similar as possible, as quantified by the relative Shannon entropy or, equivalently, by the Kullback−Leibler divergence

$$D_{KL}(p_1|p_0) = \int p_1(\mathbf{R}) \ln \frac{p_1(\mathbf{R})}{p_0(\mathbf{R})} d\mathbf{R} \quad (3)$$

Clearly, restraining a single simulation with a harmonic restraint to the data (eq 1) violates the maximum entropy principle because the restraint modifies not only the ensemble average, which is supported by the data, but also the variance of the ensemble, which is typically not supported by the ensemble-averaged data.[65,67] Hence, alternative coupling schemes have been proposed for biasing the conformational sampling by the data while applying only a minimal bias.[48,50,67,68] Pitera and Chodera showed that replacing the harmonic restraint in eq 1 with a linear restraint will generate a minimally biased ensemble. Alternatively, coupling of N parallel-replica simulations with a harmonic restraint to the data was shown to follow the maximum entropy principle in the limit of a large number of replicas.[48,68] Accordingly, the back-calculated scattering signal is first averaged among the parallel replicas

$$\bar{I}_c(q_i, \mathbf{R}_1, ..., \mathbf{R}_N) = N^{-1} \sum_{\alpha=1}^{N} I_c(q_i, \mathbf{R}_\alpha) \quad (4)$$

where α is the replica index, and henceforth coupled to the experimental data with the restraint

$$E_{exp}(\mathbf{R}_1, ..., \mathbf{R}_N; I_{exp}) = \frac{k_r N^a k_B T}{n_q} \sum_{i=1}^{n_q} \frac{[\bar{I}_c(q_i, \mathbf{R}_1, ..., \mathbf{R}_N) - I_{exp}(q_i)]^2}{\sigma_i^2} \quad (5)$$

Following Hummer and Köfinger,[36] setting the exponent a in eq 5 to $a = 1$ ensures that the optimal ensemble is recovered as $N \to \infty$. Intuitively, this decision leads to a cancelation of N in eq 5 with N^{-1} in eq 4 when computing the mean experiment-derived biasing force on atom j during an MD simulation, as given via the gradients $-\partial E_{exp}/\partial \mathbf{R}_{\alpha,j}$, where $\mathbf{R}_{\alpha,j}$ is the position of atom j in replica α. The force constant k_r is an empirical parameter that expresses the degree of confidence in the experimental data versus the (unbiased) force field. For $k_r = 0$, the unbiased ensemble is recovered; for large k_r, the ensemble reproduces the experimental data with increasing precision. For practical applications, a reasonable choice for N and k_r may be found by plotting D_{KL} and the residuals χ^2 between experimental and calculated data versus increasing N and k_r; then, the smallest values for N and k_r may be chosen that give satisfactory χ^2 and D_{KL}.[21]

In this work, we developed a method for refining MD simulations of flexible biomolecules on-the-fly to SAXS data using a parallel-replica ensemble restraint. We demonstrate the method by refining ensembles of the disordered RS peptide against experimental SAXS data. SAXS curves were computed with explicit-solvent methods developed previously that take into account accurate atomic models for both the hydration layer and for the excluded solvent.[69−71] In contrast to implicit-solvent SAXS calculations, explicit-solvent calculations do not require free fitting parameters for the hydration layer and excluded solvent.[61] We systematically investigated the influence of the force field, the number N of parallel replicas, and the force constant k_r on the refined ensembles.

COMPUTATIONAL DETAILS

Jensen−Shannon Divergence. We used the Jensen−Shannon divergence to quantify the similarity between the unbiased ensemble p_0 and the biased ensemble p_1.[72] It is defined as

$$D_{JS}(p_1|p_0) = \frac{1}{2}(D_{KL}(p_0|M) + D_{KL}(p_1|M)) \quad (6)$$

where $M = (p_0 + p_1)/2$ is the average of the two distributions and D_{KL} denotes the Kullback−Leibler divergence. The Jensen−Shannon divergence may be considered as a symmetrized and smoothed variant of the Kullback−Leibler divergence (eq 3). We used D_{JS} because it is numerically more robust than D_{KL} in the case of poor sampling. For instance, if

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

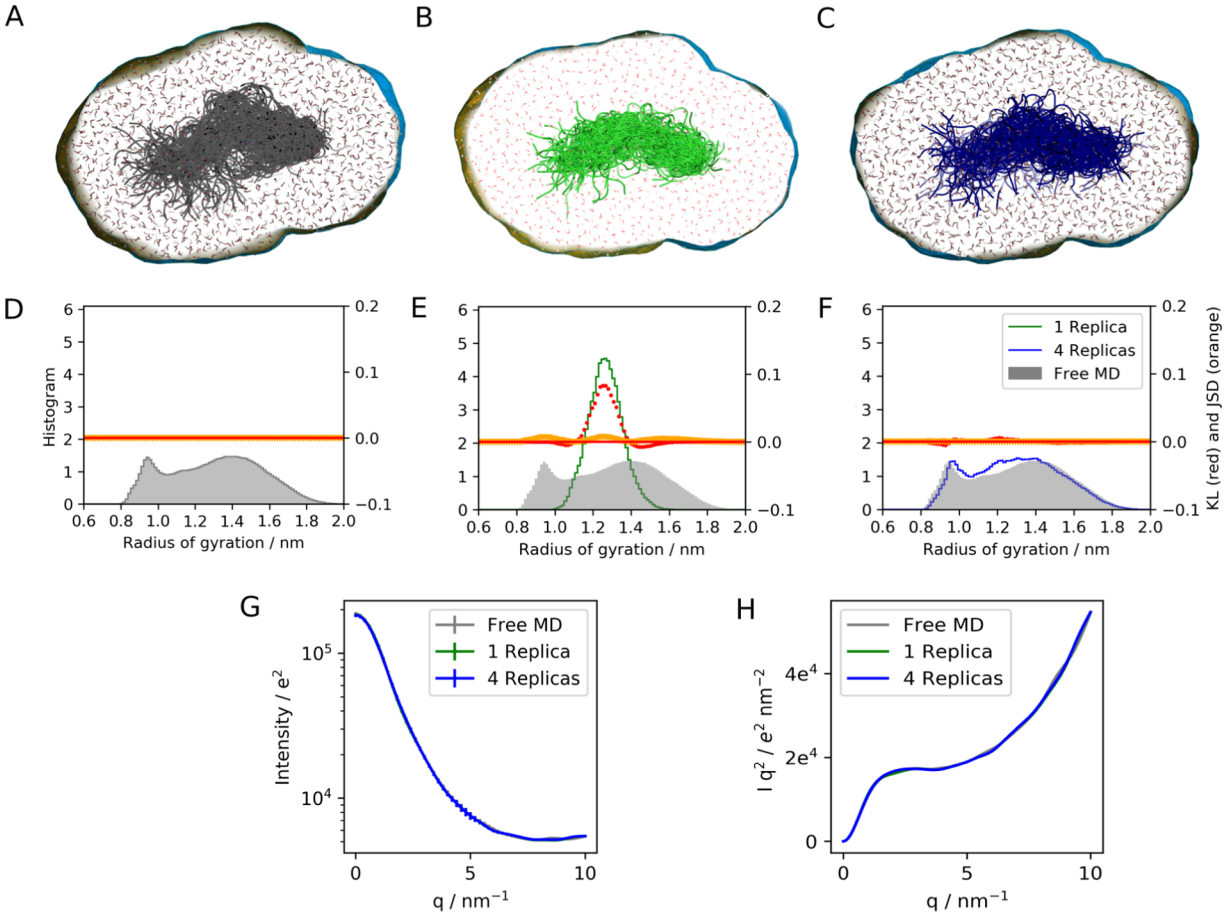


Figure 1. (A–C) Overlay of snapshots of the RS peptide of the ensembles of the (A) unbiased simulation, (B) single-replica refinement, and (C) four-replica refinement. Explicit water molecules (red/white sticks) inside the envelope (blue surface) were taken into account for computing the SAXS curves. (D–F) Distributions of radii of gyration in the ensembles: unbiased ensemble for reference (gray, D–F), biased ensemble from single-replica refinement (green, E), and biased ensemble from four-replica refinement (blue, F). The difference between the biased and unbiased ensembles is shown by the contributions to the Kullback–Leibler divergence (red) and the Jensen–Shannon divergence (orange). (G) Computed SAXS curves and (H) Kratky plots of free/unbiased MD (gray), single-replica refinement (green), and four-replica refinement (blue) are hardly distinguishable.

certain states were rarely visited in the unbiased MD simulation but are present in the refined ensemble, D_{KL} would be subject to high uncertainty or even be undefined. D_{JS} does not face such problems. Below, we report $D_{JS}(p_1^g|p_0^g)$ for the distributions of the radii of gyration $p_0^g(R_g)$ and $p_1^g(R_g)$ of the unbiased and biased ensembles, respectively.

SAXS Predictions, Coupling Protocol, and Memory Time. For each of the parallel replicas, SAXS curves were computed on-the-fly using explicit-solvent predictions, similar to previous work.[59,69] Briefly, the intensity of replica α at simulation time t was computed as

$$I_{c,\alpha}(q, t; \tau) = \langle D_\alpha(\mathbf{q}, t; \tau)^{(\omega)} \rangle \Omega \tag{7}$$

where $\langle\cdot\rangle_\Omega$ denotes the orientational average, τ is a memory time[59] (see below), and

$$D_\alpha(\mathbf{q}, t; \tau)^{(\omega)} = \langle |\tilde{A}_\alpha(\mathbf{q}, t)|2 \rangle_{t;\tau}^{(\omega)} - \langle |\tilde{B}(\mathbf{q}, t)|^2 \rangle^{(\omega)} - 2\mathrm{Re}[\langle \tilde{B}(\mathbf{q}, t) \rangle^{(\omega)} [\langle \tilde{A}_\alpha(\mathbf{q}, t) \rangle_{t;\tau}^{(\omega)} - \langle \tilde{B}(\mathbf{q}, t) \rangle^{(\omega)}]] \tag{8}$$

is the buffer-subtracted scattering intensity at fixed solute orientation ω. $\tilde{A}_\alpha$ and $\tilde{B}$ denote the instantaneous scattering amplitudes of the peptide in replica α including the hydration layer and of (the same volume of) purely water, respectively. Re[·] is the real part. The amount of explicit water included in the calculation of $\tilde{A}_\alpha$ and $\tilde{B}$ was defined by a spatial envelope that was constructed at a distance of approximately 9 Å from all peptides atoms, resulting in a volume of 103 nm^3 (Figure 1A–C). The same envelope was used throughout this study. The symbol $\langle\cdot\rangle$ denotes a uniformly weighted temporal average. The symbol $\langle\cdot\rangle_{t;\tau}$ denotes a moving average with a memory kernel that decays exponentially into the past with time constant τ

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

where $\mathcal{N}(t) = \int_0^t \mathrm{e}^{(t'-t)/\tau} \, \mathrm{d}t'$ is a normalization constant. The parameter τ is specified prior to the simulation. Subsequently, the intensities $I_{c,\alpha}$ were averaged over the N replicas (eq 4) and the averaged intensities were coupled to the experimental data I_{exp} with a harmonic restraint (eq 5). Hence, the calculated SAXS curves that are compared with the experiment represent not only an average over the replicas but also a time average with memory time τ. As shown in the results, this leads to some heterogeneity in the refined ensembles even if only a single replica is coupled to the data. For more details on the

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103–5115

explicit-solvent SAXS predictions, we refer to previous work.[59,69]

The value of τ should be chosen large enough to ensure that the buffer subtraction is converged over a time interval of $1-2\tau$. We carefully assessed the convergence of the intensities and found that, for the RS peptide studied here, a value $\tau = 100$ ps is reasonable given that the forward scattering intensity is fixed to a preset value (see below). Further, we tested the effect of τ on the ensemble and found that modulating τ between 50 and 200 ps has only a small effect on the refined ensembles.

The time integral in eq 9 was approximated by a discrete sum, which was updated using simulation frames every 0.5 ps.[59] The orientational average was computed numerically using 342 **q**-vectors per absolute value of q. A total of 1000 frames of a water box were used to compute the pure-water scattering contributions $\langle\tilde{B}\rangle$ and $\langle|\tilde{B}|^2\rangle$ in eq 8. The force constant k_r was set to 1 in this work, and the scaling exponent a for the restraining energy was chosen between 0 and 2 (see below). The variances of the intensities $\sigma_i^2 = \sigma_{e,i}^2 + \sigma_{c,i}^2 + \sigma_{buf,i}^2$ were computed from statistical experimental errors σ_e, statistical calculated errors σ_c, and from systematic errors σ_{buf}, which were modeled via an uncertainty of the buffer density of 0.1%.[59,69] Calculated errors σ_c were computed assuming (i) that frames every 0.5 ps have statistically independent solvent conformations and (ii) using Gaussian error propagation. In the control simulations that were coupled to a precomputed SAXS curve (see below; Figure 1), $\sigma_{e,i}$ was taken as the statistical error of the precomputed curve. Gradients of the SAXS intensity, as required to compute SAXS-derived forces, were computed as described previously.[59]

Matching the Absolute Scale of SAXS Curves and Fixing the Forward Scattering during Refinement. Since an IDP (including its hydration layer) imposes only a small contrast relative to the pure solvent, the explicit-solvent SAXS calculations required tens of nanoseconds of simulation time to converge the calculated forward scattering. To ensure that the calculated SAXS curve was reasonably converged within memory time τ, as required for coupling the simulations on-the-fly to the data, we first computed a SAXS curve I_c^{prelim} from several SAXS-restrained simulations with four replicas for each force field. Then, the absolute scale of I_{exp} was adjusted such that the forward scattering (at $q = 0$) matched between I_c^{prelim} and I_{exp}. Accordingly, we obtained for the forward scattering 186 828 e^2, 208 872 e^2, and 208 415 e^2 for simulations with CHARMM36m/cTIP3P,[11] Amber99SBws/TIP3P,[58,73] and Amber99SBws/TIP4P-D,[58,74] respectively. Here, cTIP3P denotes the CHARMM-modified TIP3P model with Lennard-Jones interactions on hydrogen atoms. Henceforth, during production SAXS-restrained simulations, we fixed the calculated forward scattering $I_c(q = 0)$ to the value $I_c^{prelim}(q = 0)$ obtained from the preliminary simulations. This was achieved by adding a small uniform electron density $\delta\rho_B$ to the pure-solvent system, which translates into a correction $\delta\tilde{B}(\mathbf{q}) = \delta\rho_B\tilde{\Theta}_e(\mathbf{q})$ for the pure-solvent scattering amplitude. Here, $\tilde{\Theta}_e(\mathbf{q})$ denotes the Fourier transform of a unit density in the envelope. We found that fixing $I_c(q = 0)$ during refinement calculations greatly accelerates the convergence of the SAXS curve.

Simulation Details. The SAXS calculations and restraints were implemented into an in-house modification of GROMACS 4.6.[75,76] This software is available from the authors upon request. Hydrogen atoms of the proteins were modeled as virtual interaction sites allowing an integration time step of 4 fs. Electrostatic interactions were computed with the particle-mesh Ewald scheme.[77,78] A cutoff at 1 nm was applied to the direct-space Coulomb interactions and at 0.8 nm to the Lennard-Jones interactions. The short cutoff was chosen to reduce the considerable computational cost of the simulations presented here, and it was justified by the fact that we focused on biasing ensembles with experimental data and not on deriving highly precise unbiased ensembles. The Lennard-Jones potential was shifted to zero at the cutoff. The bond lengths and angles of water molecules were constrained with the SETTLE algorithm,[79] and all other bonds were constrained with LINCS.[80] The pressure was set to 1 bar using the Berendsen barostat ($\tau = 4$ ps).[81] During the simulations, the temperature was controlled at 300 K using a tight stochastic Langevin dynamics integration scheme ($\tau = 0.3$ ps), motivated from the fact that SAXS-driven MD simulations are not strictly energy-conservative.[82] The simulation parameters were identical in free and SAXS-driven simulations. Starting structures for the replica simulations were taken from different time frames of an unbiased MD simulation of 400 ns. SAXS-restrained simulations were carried out for 400 ns per replica. Simulation frames were saved every 2 ps for later analysis. Secondary structure content was computed with DSSP, version 2.0.4.[83] The convergence of the ensembles was assessed by binning analysis and by computing D_{JS} as a function of simulation time (Figures S1 and S2).

RESULTS AND DISCUSSION

The RS peptide with the sequence GAMGPSYG$(RS)_8$ is a well-characterized IDP subject to multiple previous studies. For instance, this peptide has been used as a benchmark system to test the accuracy of several force fields against experimental SAXS and NMR data.[10] It was also included in a SAXS and NMR data set of 15 peptides and 20 proteins used to optimize the CHARMM36m force field.[11] Hence, CHARMM36m provides a suitable model for the RS peptide and, as such, provides a starting point to validate our method.

Control: Ensemble Refinement Against a SAXS Curve Computed from a Known Ensemble. As a control, we first tested whether ensemble refinement of the RS peptide is capable of recovering a given unbiased ensemble. To this end, we carried out free MD simulations of the RS peptide with CHARMM36m with a total simulation time of 8 μs such that the distribution of the radius of gyration was reasonably converged. We computed the SAXS curve from the converged ensemble and, henceforth, coupled simulations to the computed SAXS curve. In this case, ensemble-refined simulations are expected to recover the unbiased ensemble. Any difference between the unbiased and the refined ensemble would either indicate a violation of the maximum entropy principle or reveal another unphysical bias in the calculations.

The RS peptide adopts a heterogeneous ensemble during free, unbiased simulations with CHARMM36m/cTIP3P, as visualized by an overlay of MD frames (Figure 1A) and quantified by the wide distribution of the radius of gyration R_g computed from the peptide atoms of all MD frames (Figure 1D, gray histogram). The SAXS curve computed from the free simulation (Figure 1G, gray) is rather featureless, as is common for SAXS curves of IDPs. Upon coupling a single replica to the computed SAXS curve, an overly restrained ensemble is generated, as is evident from the R_g distribution that is far narrower than the distribution of the true underlying ensemble (Figure 1E, green, compared to gray distribution)

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

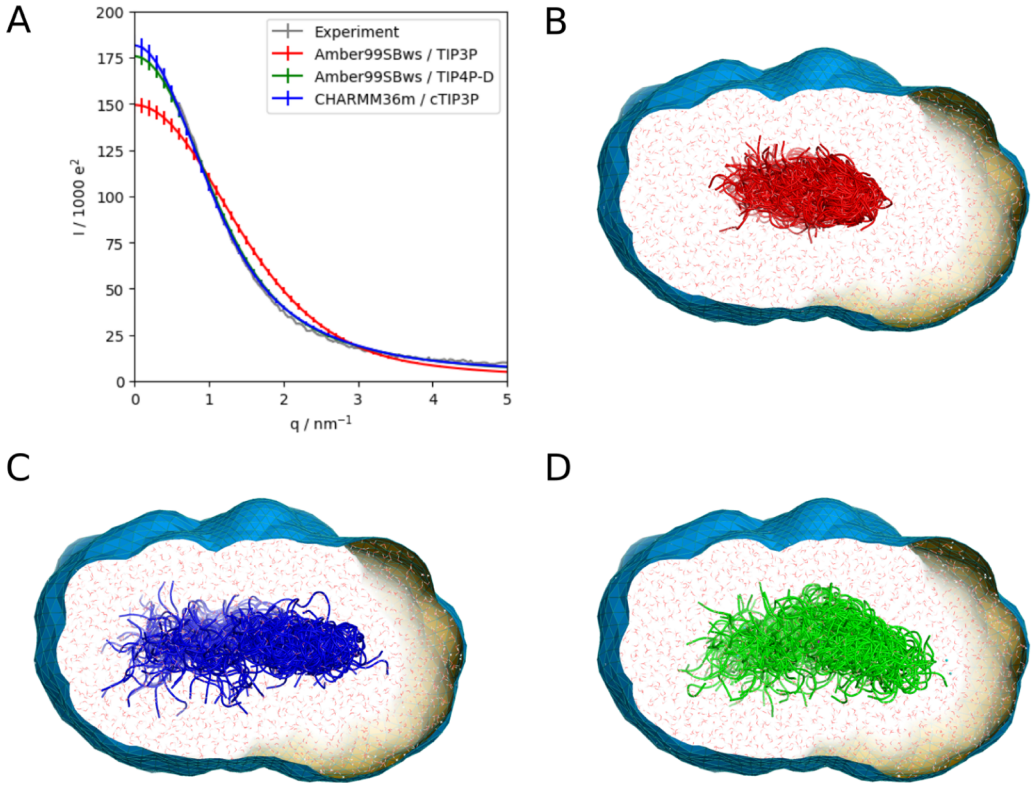


Figure 2. (A) SAXS curves calculated from free MD simulations with Amber99SBws/TIP3P (red), Amber99SBws/TIP4P-D (green), and CHARMM36m/cTIP3P (blue). The experimental curve taken from ref 10 is shown in gray. (B–D) Overlay of snapshots from free MD simulations, using color coding according to the curves in (A).

and from the relatively well-defined, homogeneous snapshots shown as overlay of MD frames in Figure 1B. The large discrepancy between the unbiased ensemble and the single-replica-refined ensemble is further quantified in Figure 1E, presenting the contributions of different R_g values to the Jensen–Shannon divergence D_{JS} (orange curve) and to the Kullback–Leibler divergence D_{KL} (red curve). Notably, contributions to D_{KL} along R_g may be partly negative, whereas the overall D_{KL} integrated over R_g is always positive. Overall, as expected, the analysis demonstrates that coupling a single replica to the SAXS curve leads to a far too narrow ensemble, violating the maximum entropy principle.

In contrast to single-replica refinement, coupling four parallel replicas of the RS peptide leads to an ensemble in favorable agreement with the unbiased ensemble, as evident from visual inspections of the trajectories and from the R_g distribution (Figure 1C,F). Hence, the four-replica refinement maintains the heterogeneity of the ensemble in agreement with the maximum entropy principle. This finding is confirmed by the Jensen–Shannon divergence D_{JS} between the unbiased and refined R_g distributions: for single-replica and four-replica refinement, we find D_{JS} values of 0.216 and 0.005, respectively, demonstrating a greatly improved similarity between the four-replica ensemble and the unbiased ensemble compared to single-replica ensemble. The agreement between the unbiased and four-replica-refined ensembles serves as a first validation for the method proposed here.

Figure 1G,H presents the SAXS curves and Kratky plots computed from the free and biased simulations. Notably, the curves are nearly indistinguishable among the three ensembles, despite the fact that the ensemble generated from the single-replica refinement is far narrower than the ensembles from the free MD or from four-replica refinement. This demonstrates that the ensemble-averaged SAXS curve does not encode the heterogeneity of the ensemble and, consequently, SAXS curves from a heterogeneous ensemble may be easily misinterpreted by fitting an overly narrow ensemble to the SAXS curve. In contrast, ensemble refinement with commitment to the maximum entropy principle allows for the maintenance of a correct ensemble width.

Identifying the Minimum Number of Replicas *N* for Ensemble Refinement. Having validated that multireplica refinement is capable of preserving the underlying ensemble generated by the force field (previous paragraph), we turn toward a more challenging task, namely, ensemble refinement against experimental data using a force field that leads to an incorrect ensemble in an unbiased simulation. Figure 2A compares the experimental SAXS curve of the RS peptide with SAXS curves from free MD simulations using three different combinations of protein and water force fields: Amber99SBws/TIP3P[58,73] (red), Amer99SBws/TIP4P-D[58,74] (green), and CHARMM36m/cTIP3P[11,84] (blue). Evidently, Amer99SBws/TIP4P-D and CHARMM36m/cTIP3P yield good agreement with the experiment, reflecting that these force fields were optimized for modeling IDPs. In contrast, Amber99SBws/TIP3P leads to an overly collapsed ensemble,[58] as evident from the MD snapshots (Figure 2B) and from a SAXS curve that decays too slowly along q (Figure 2A, red). Hence, for simulations of the RS peptide with Amber99SBws/TIP3P, experimental data are required to obtain a reasonably correct ensemble, providing a realistic test case for our method.

Roux and Weare as well as Cavalli et al. showed that the correct ensemble is obtained in the limit of a large number of replicas, $N \rightarrow \infty$.[48,68] The number of replicas required in

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103–5115

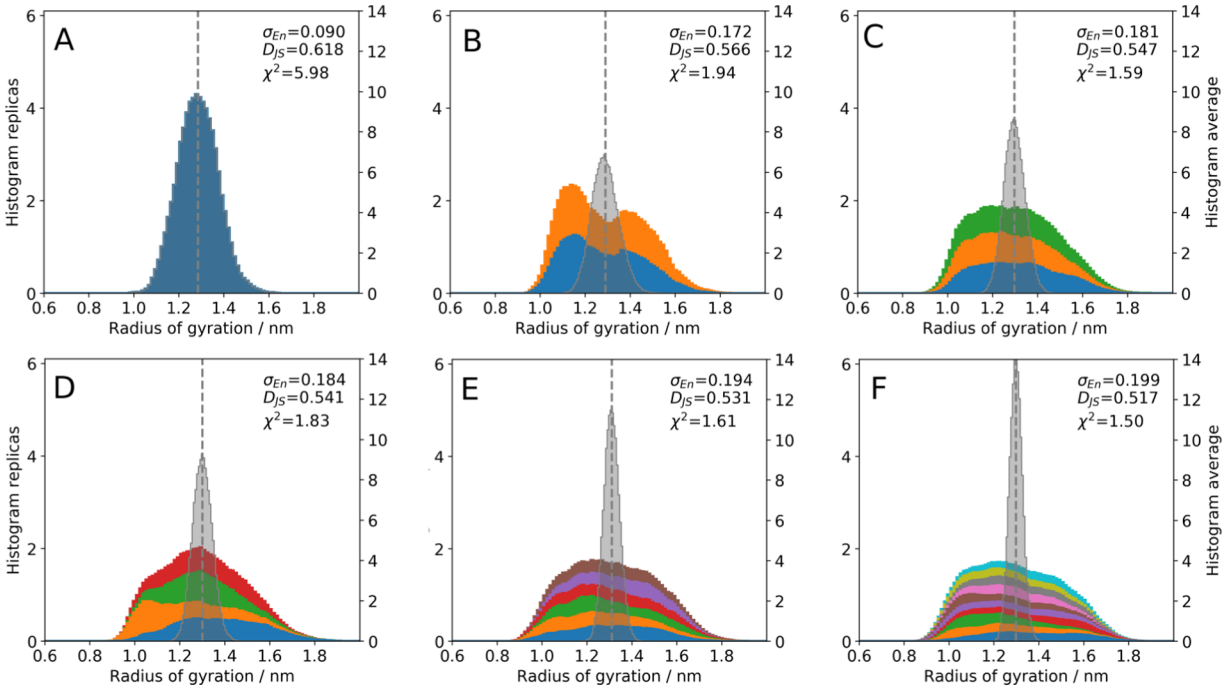


Figure 3. Distributions of the radius of gyration R_g in ensembles refined by restraining 1, 2, 3, 4, 6, or 10 replicas (A–F) to the experimental SAXS data of the RS peptide, simulated with Amber99SBws/TIP3P. Colored stacked histograms show R_g distributions of the individual replicas. Gray histograms show distributions of R_g averaged over the replicas at each time step. The gray dashed lines indicate the average over the simulation time. The standard deviation σ_{En} of the ensemble, Jensen–Shannon divergence D_{JS}, and χ^2 between calculated and experimental SAXS curves are shown as inset for each ensemble.

practice to obtain a reasonably accurate ensemble may depend on the system. To test how many replicas are required for the RS peptide, we restrained 1, 2, 3, 4, 6, or 10 replicas to the experimental SAXS curves (Figure 3). For these simulations, the force constant was scaled linearly with the number of replicas, as implied by eq 5 using the exponent $a = 1$. We first validated that the simulations are capable of restraining the average R_g at the expected value, as evident from the fact that the distributions of R_g averaged over the replicas at each time step peak near the experimental value of 1.29 nm (Figure 3A, blue histogram, and B–F, gray histograms; compare with gray dashed lines). Notably, the replica-averaged R_g distributions become narrower with increasing N reflecting a tighter coupling of the replica average to data, presumably because memory effects owing to eq 9 are mitigated by averaging over an increasing number of replicas (Figure 3, gray histograms). Next, the restrained ensembles were evaluated by computing (i) the Jensen–Shannon divergence $D_{JS}(p_1^g|p_0^g)$ between the unbiased and the refined ensembles, (ii) the standard deviation σ_{En} of the R_g distribution, and (iii) χ^2 between the calculated and experimental SAXS curves (Figure 4).

With increasing number of replicas N, D_{JS} decreases, demonstrating an increasing similarity between the refined and the unbiased ensembles and, consequently, demonstrating that the SAXS-derived bias is indeed increasingly "minimal" (Figure 4A). The decreasing D_{JS} is in line with Roux, Cavalli, and co-workers[48,68] in the sense that, with increasing N, the refined ensemble approaches the desired, minimally biased ensemble. This finding is further corroborated by the increasing standard deviation σ_{En} of the ensemble (Figure 4B), implying that the entropy of the ensemble increases with N. To quantify the agreement between the refined ensemble and the data, Figure 4C presents χ^2 versus N. For the

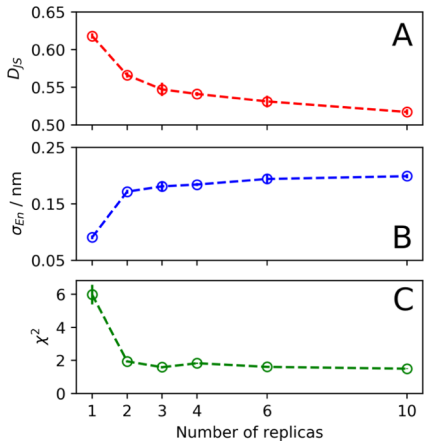


Figure 4. Convergence of the restrained ensemble with increasing number of replicas. (A) Jensen–Shannon divergence, (B) standard deviation of the R_g distribution as a measure of the width of the refined ensemble, and (C) reduced χ^2 between experimental and calculated SAXS curve. The vertical bars indicate standard errors computed by binning the simulations into blocks of 100 ns (hardly visible for some points).

Amber99SBws/TIP3P system, the agreement between the calculated and the experimental data improves significantly between $N = 1$ and $N = 2$ but hardly for $N > 2$, in contrast to D_{JS} that strongly decreases between $N = 1$ and $N = 3$ and henceforth slightly decreases up to the case $N = 10$ (Figure 4, compare A with C). This finding confirms the notion that (i) a low χ^2 does by far not imply that the ensemble is correct[21] and (ii) analyzing χ^2 alone would be insufficient for finding a good choice for N; instead, both D_{JS} and χ^2 should be considered to find a justified choice for N. For the simulations shown below,

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103–5115

we used $N = 4$, providing a reasonable balance between low D_{JS} and computational simplicity.

Previous studies restrained many more replicas for refining an ensemble against experimental data, in the order of tens of replicas.[48,55] The large number or replicas were used because the back-calculated data $\bar{I}_c$, which are compared with the experimental data I_{exp} in eq 5, was typically computed using purely the current simulation frame. Consequently, in such previous studies, many parallel replicas were required to represent the overall heterogeneity of the ensemble at each simulation time. The rapid convergence of the ensemble with fewer replicas, as found here, may be rationalized by our method for computing the SAXS curve. Namely, owing the time averaging with memory time τ (eq 9), SAXS curves computed for individual replicas already account for fluctuations occurring on the time scale of approximately 2τ (100−200 ps), covering side-chain fluctuations and rapid small-scale fluctuations of the peptide backbone. Consequently, to make our back-calculated SAXS curve represent the overall heterogeneity covering both rapid small-scale and slow large-scale fluctuations, only few replicas are required.

Force Constant and Scaling of the Restraining Energy with the Number of Replicas *N*. The prefactor $k_r N^a$ of E_{exp} (eq 5) specifies the weight of the experimental bias compared to the force-field energy E_{MD}, thereby quantifying our confidence in the data relative to prior knowledge.[85] Both the force constant k_r and the scaling exponent a modulate the weight of E_{exp}. However, because the uncertainty of the force field as well as systematic errors in the experimental data are in practice unclear, the force constant is often treated as a heuristic parameter. It was argued that a good choice for the force constant is obtained by plotting χ^2 versus the force constant k_r and by selecting the smallest k_r that leads to a satisfactory agreement with the data. Apart from a suitable choice for k_r with a fixed number of replicas N, the suitable exponent of a to scale the restraining energy with N has been discussed.[36,65,68]

To test the influence of the scaling, we conducted several ensemble refinement simulations with four replicas and with an increasing scaling exponent a of 0, 1, 1.5, and 2 (Figure 5). As expected, we find that increasing a leads to slightly decreasing χ^2, reflecting an increasing weight on the experimental data. Overall, however, the restraining energy has only a minor effect on the refined ensembles. A value of $a = 1$, scaling the restraining energy linearly with the number of replicas, is appropriate for the system considered here.

Small Influence of the Force Field on Refined Ensembles. It is well established that IDP ensembles generated by free MD simulations strongly depend on the force field.[10,74] For instance, Amber99SBws/TIP3P favors an overly collapsed RS peptide, whereas Amber99SBws/TIP4P-D and CHARMM36m/cTIP3P yield RS ensembles in reasonable agreement with SAXS data, as evident from the average radii of gyration in the free simulations (Table 1), from the calculated SAXS curves (Figure 2A), and from the MD snapshots (Figure 2B−D). This prompted us to test whether SAXS-restrained ensembles are likewise force field-dependent or whether the experimental data are capable of mitigating the imperfections of force fields.

Figure 6 compares the generated ensembles from free and SAXS-restrained simulations with CHARMM36m/cTIP3P (A−C), Amber99SBws/TIP3P (D−F), and Amber99SBws/TIP4P-D (G−I). For reference, Figure 6A/D/G present the R_g distributions from free simulations (gray-shaded areas), confirming that the R_g distributions are strongly force-field-dependent, in line with the force-field-dependent conformations and SAXS curves shown in Figure 2. Namely, in free simulations, Amber99SBws/TIP3P leads to an overly collapsed ensemble restricted to low R_g values (Figure 6D, gray), whereas Amber99SBws/TIP4P-D and CHARMM36m/cTIP3P lead to more expanded ensembles (Figure 6A/G, gray). Upon restraining only a single replica to the SAXS curve, for all three force fields, the ensemble-averaged radius of gyration agrees with the experiment, but all ensembles become too narrow (Figure 6B/E/H, blue histograms; A/D/G, green histograms, and Table 1). The overly narrow ensembles are evident from narrow R_g distributions and from an increased Jensen−Shannon divergence between the unbiased and biased R_g distributions (see insets in Figure 6). This demonstrates that the experiment-derived bias is not minimal during single-replica refinement.

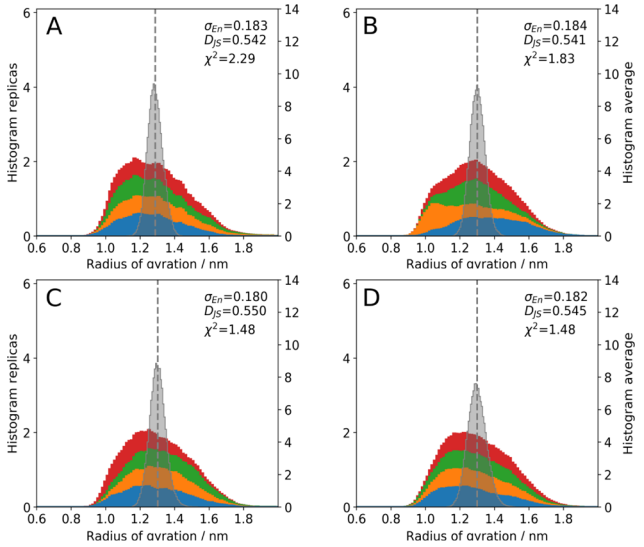


Figure 5. SAXS-restrained ensembles generated with four parallel replicas ($N = 4$) and Amber99SBws/TIP3P, and using an increasing scaling exponent a for the number of replicas N ((A) $a = 0$, (B) $a = 1$, (C) $a = 1.5$, (D) $a = 2$). Colored stacked histograms show R_g distributions of the individual replicas. The gray histograms show distributions of R_g averaged at each time step over the four replicas.

In contrast, upon restraining four replicas using any of the force fields, the R_g distributions become wider than the single-replica ensembles while maintaining agreement with the experimental average. The Jensen−Shannon divergence also strongly decreases, demonstrating that the resulting ensembles are more similar to the unbiased simulation (Figure 6C/F/I, D_{JS} labels in subplots). As a key finding of this study, the R_g distributions of the refined four-replica ensembles are similar among the three force fields. This suggests that the SAXS-derived restraints are capable of mitigating major imperfections in the force fields, leading to ensembles that are much less force-field-dependent compared to ensembles from free simulations. Notably, this finding is most obvious for the Amber99SBws/TIP3P force field, which yields overly collapsed ensembles for the RS peptide in free simulations (Figure 6D, gray), but an ensemble similar to results from the optimized CHARMM36m/cTIP3P or Amber99SBws/TIP4P-D when coupled to SAXS data (Figure 6C/F/I).

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

Table 1. Mean Radius of Gyration R_g in nm with Several Force Fields, Taken from the R_g Distributions in Figure 6 (Experiment: 1.290 nm[10])[a]

		SAXS-restrained MD	
force field	free MD	$N = 1$	$N = 4$
CHARMM36m/cTIP3P	1.30 ± 0.02	1.27 ± 0.03	1.280 ± 0.003
Amber99SBws/TIP3P	0.88 ± 0.03	1.285 ± 0.004	1.301 ± 0.002
Amber99SBws/TIP4P-D	1.11 ± 0.02	1.29 ± 0.01	1.29 ± 0.01

[a]Error estimates were calculated using binning analysis.[86]

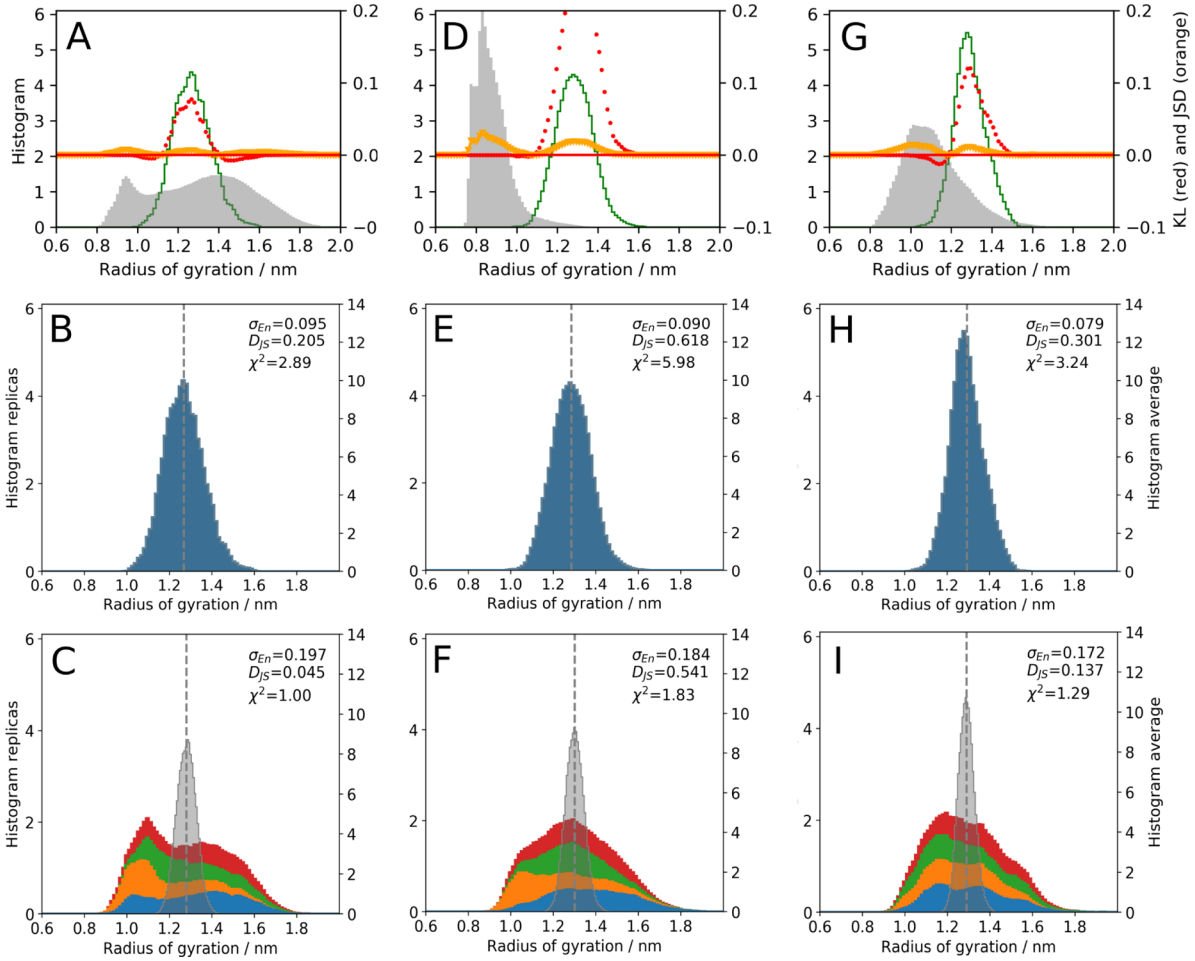


Figure 6. Distributions of the radius of gyration R_g from single-replica refinement (top and middle row) with the multireplica refinement (bottom row, four replicas) for three different force fields and water models: CHARMM36m/cTIP3P (A−C), Amber99SBws/TIP3P (D−F), Amber99SBws/TIP4P-D (G−I). For reference, distributions from the free MD simulation are shown as gray histograms (top row). Distributions from SAXS-restrained simulations as color-filled histograms. Distributions of individual replicas are shown as stacked histograms. Width of the R_g distribution σ_{En}, the Jensen−Shannon divergence D_{JS}, and the reduced χ^2 between calculated an experimental SAXS curves are listed in the subplots. (A/D/G) The red and orange curves illustrate contributions to the Kullback−Leibler and Jensen−Shannon divergence along R_g, respectively, quantifying the difference between the distributions from free (gray histograms) and single-replica simulations (green lines). (C/F/I) Colored stacked histograms show R_g distributions of the individual replicas. The gray histograms show distributions of R_g averaged at each time step over the four replicas.

CROSS-VALIDATION OF SAXS-RESTRAINED ENSEMBLES

SAXS-restrained ensembles generated with different force fields are similar to those judged by their R_g distributions (Figure 6). However, this does not exclude the possibility that other observables differ among the ensembles. Hence, to further compare the ensembles, we analyzed $^3J_{HN\text{-}H\alpha}$ couplings, distance matrices between the C_α atoms, and the secondary structure content.

SAXS curves and the R_g distributions analyzed above provide structural information averaged over all interatomic distances. As a spatially more detailed characterization, to test the effect of SAXS restraints also on individual interatomic distances, we computed C_α−C_α distance matrices from free and SAXS-refined simulations. Figure 7 shows the differences of C_α−C_α distance matrices between simulations with Amber99SBws/TIP3P and CHARMM36m/cTIP3P. As expected, the unrestrained, overly collapsed Amber99SBws/TIP3P ensemble strongly differs from the more extended

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

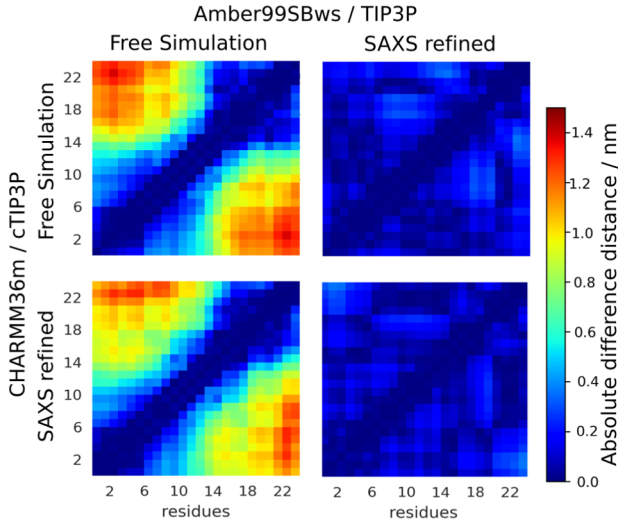


Figure 7. Distance heat map comparing the mean distances between pairs of C_α atoms between free simulations and four-replica-restrained simulations, using either CHARMM36m/cTIP3P or Amber99SBws/TIP3P. Color coded is the absolute value of the difference of C_α–C_α distances: $d_{ij} = |\langle r_{ij}\rangle_{Sim1} - \langle r_{ij}\rangle_{Sim2}|$, where $\langle r_{ij}\rangle$ denotes the mean distance of the C_α atoms of residues i and j. The analysis demonstrates that C_α distances in free Amber simulations greatly differ from free or refined CHARMM simulation (left column). In contrast, C_α distances in refined Amber simulations are similar to free or refined CHARMM simulations (right column).

CHARMM36m/cTIP3P ensemble (Figure 7, top left). However, applying a SAXS restraint to the Amber99SBws/TIP3P force field strongly reduces the differences to the free CHARMM36m/cTIP3P simulations (Figure 7, top right), confirming that the SAXS restraint mitigates major imperfections of Amber99SBws/TIP3P. In addition, the C_α distances are similar among the restrained Amber99SBws/TIP3P and CHARMM36m/cTIP3P simulations (Figure 7, bottom right), confirming that the SAXS-restrained ensembles of the RS peptide are similar among different force fields.

While SAXS data are sensitive with respect to the overall shape of the ensemble, $^3J_{HN\text{-}H\alpha}$ coupling data probe the backbone ϕ dihedrals and thus provide complementary local structural information. ^3J-coupling constants computed with the Karplus relation are presented in Figure 8.[87] Both free and SAXS-restrained CHARMM36m/cTIP3P simulations reveal

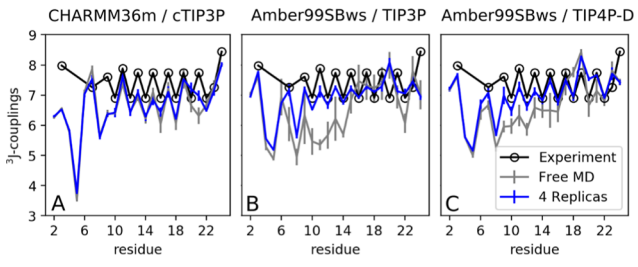


Figure 8. $^3J_{HN\text{-}H\alpha}$-coupling data from experiment (black circles), calculated from free MD simulations (gray), and from SAXS-restrained ensembles with four replicas (blue). Experimental data from ref 10; the error bars denote standard errors over four replicas. Upon coupling to SAXS data, the RMSD between calculated and experimental ^3J-coupling decrease from 0.76 to 0.67 (Charmm36m/cTIP3P), from 1.14 to 0.58 (Amber99SBws/TIP3P) and from 0.88 to 0.45 (Amber99SBws/TIP4P-D), respectively.

reasonable agreement with experimental ^3J-couplings taken from ref 10 (Figure 8A). This finding is expected because CHARMM36m/cTIP3P has previously been validated against experimental data of the RS peptide.[11] In contrast, free Amber99SBws/TIP3P and Amber99SBws/TIP4P-D simulations reveal considerable deviations with respect to experimental ^3J-couplings, mainly between residue numbers 6 and 14, possibly owing to imperfections in the backbone dihedrals in the Amber99SBws force field (Figure 8B,C, gray and black lines). However, the agreement between calculated and experimental ^3J-couplings greatly improves upon introducing the SAXS restraint (Figure 8B,C, blue lines). The RMSD between calculated and experimental ^3J-couplings decreases from 1.14 to 0.58 and from 0.88 to 0.45 for Amber99SBws/TIP3P and Amber99SBws/TIP4P-D, respectively. A typical example is given by residue 9, for which coupling Amber99SBws/TIP4P-D simulations to SAXS data triggers an increased population of the backbone conformation with $\phi \approx 60°$ and $\phi \approx -135°$, thereby achieving improved agreement with experimental ^3J-couplings (Figures S5 and 8C). The improved agreement to ^3J-couplings is remarkable considering that the incorporated SAXS data are not sensitive to short-distance information encoded in the ^3J-couplings. Hence, the improvement of NMR data demonstrates that our replica-restrained MD simulations do not overfit the SAXS data, but instead yield conformational ensembles with improved structure at both the global and local levels.

Further, a comparison of ^3J-couplings between different force fields (Figure S3) demonstrates that the backbone structure of SAXS-restrained ensembles is overall less force-field-dependent compared to the backbone structure in free simulations. An exception is the N-terminal region of residue numbers ≤5, where a major difference between Amber and CHARMM simulations remains even in SAXS-refined ensembles (Figures S4 and 8C).

As an additional structural probe of peptide conformations, we analyzed the secondary structure content of the ensembles using the DSSP software (Figure 9).[83] In free simulations, Amber99SBws/TIP3P yields a larger population of α-helices

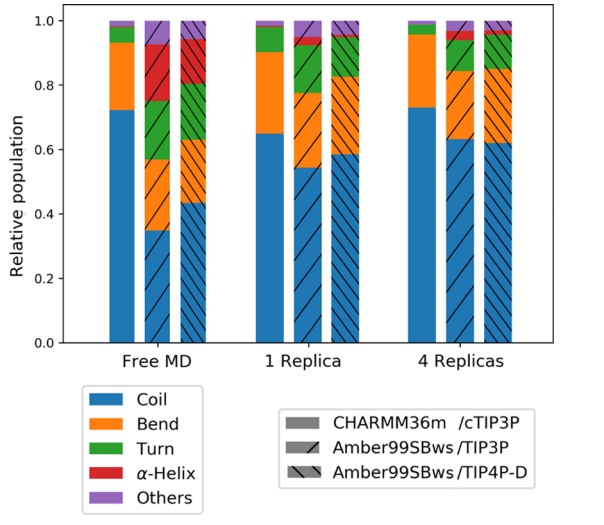


Figure 9. Secondary structure population in free simulations (left), SAXS-restrained simulations with a single replica, (middle) and with four replicas (right). Color encodes the type of secondary structure, and the pattern indicates the force field (see legends).

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

and turns compared to CHARMM36m/cTIP3P, whereas the latter yields a larger population of coils. Upon coupling the simulations to SAXS data, the secondary structure populations with Amber99SBws are shifted from α-helices and turn to coils, whereas the populations with CHARMM36m are relatively unaffected. Consequently, the secondary structure content of the SAXS-refined ensembles is more similar among different force fields compared to the content in the ensembles from free simulations.

Taken together, SAXS-restrained ensembles of the RS peptide refined from different force fields yield similar ensembles as quantified by R_g distributions, C_α distance matrices, NMR data, and secondary structure content. Hence, although residual force field effects certainly remain, by incorporating SAXS data, the influence of the force field is greatly reduced. In addition, cross-validation against NMR data suggests that the SAXS-restrained ensembles were not overfitted.

CONCLUSIONS

We presented a multireplica SAXS refinement method for generating ensembles with commitment to the maximum entropy principle. The restraint is designed to guide the simulation into agreement with experimental SAXS data while applying only a minimal bias. In contrast to established methods that reweight proposed ensembles a posteriori, the aim of this method is to integrate experimental SAXS data on-the-fly into MD simulation, leading to conformational sampling in agreement with the data. Notably, the SAXS curve predictions applied here involve explicit-solvent calculations, taking into account atomistic representations for both the hydration layer and the excluded solvent, thereby avoiding any solvent-related fitting parameters. For SAXS predictions of IDPs, however, since the contrast between solute and solvent is low, careful convergence assessments are recommended. To obtain converged on-the-fly SAXS predictions of the RS peptide studied here, it was necessary to fix the forward scattering $I(q = 0)$, where $I(q = 0)$ was taken from preliminary SAXS-restrained simulations conducted prior to the production refinement simulations.

While our method requires a small computational overhead during the simulation (~15%) compared to a simulation of the same system without the SAXS bias, it provides two main advantages compared to reweighting schemes: (a) the SAXS-guided on-the-fly conformational sampling allows for analyses that require averaging over a large number of frames, which would be tedious when using a reweighting scheme. Typical examples would be calculations of free energies, entropies, and energies, as well as calculations of slowly converging structural and kinetic observables involving rare events. Using reweighting, such analysis may require postprocessing of hundreds of thousands of frames (or even more) on the hard drive, which may become intractable. With our method, in contrast, such averages may be carried out on-the-fly during the SAXS-guided simulation; (b) if the unbiased and biased ensembles exhibit only a small overlap owing to significant force-field limitations, reweighting schemes may converge slowly and lead to refined ensembles that may be dominated by only a few frames. In other words, in such cases, the free simulation carried out prior to reweighting will spend most of the simulation time in structures that hardly contribute to the sought-after averages, which is computationally inefficient. The refinement method shown here does not face such problems.

We systematically investigated the effect of (i) the number of parallel replicas N and (ii) the scaling of the restraining energy with N. We found that the refined ensemble greatly improves when switching from a single to up to four replicas, as quantified by the Jensen−Shannon divergence. Using more than four replicas has only smaller effects on the ensembles, suggesting that four replicas provide a reasonable balance between accuracy and simplicity for the RS peptide. Notably, because (i) parallel-replica simulations scale almost linearly on parallel architectures and (ii) the replicas provide statistically independent simulations, using parallel replicas is not a limitation but instead convenient and computationally efficient for obtaining converged observables. In addition, we found that the scaling of the restraining energy with N is not critical in our case. We suggest that using a force constant of unity combined with a linear scaling with the number of replicas N is reasonable (corresponding to $k_r = 1$, $a = 1$, eq 5).

Remarkably, the applied force field has only a small effect on the refined ensembles, even when using the Amber99SBws/TIP3P force field, which is a rather poor force field for IDPs in free simulations. The refined ensembles were similar between force fields as quantified by the R_g distribution, C_α distance maps, ^3J-couplings, and secondary structure content. This agreement may indicate that Amber99SBws/TIP3P exhibits only modest imperfections, allowing one to overrule the imperfections upon adding the SAXS data. The agreement among different force fields further suggests that the multireplica refinement did not overfit the SAXS data. It will be interesting to test in a future study if our findings for the RS peptide also hold for other disordered system such as different IDPs or rigid domains connected by flexible linkers.

Overall, the present study suggests that incorporating SAXS data on-the-fly into MD simulations of disordered systems may greatly reduce a force-field bias. The method will be useful for obtaining physically precise simulations of disordered systems and for advancing our understanding of unstructured biological macromolecules.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.9b00338.

Binning analysis of the distribution of the radius of gyration (Figure S1); convergence of the Jensen−Shannon divergence D_{JS} with simulation time (Figure S2); comparison of $^3J_{HN\text{-}H\alpha}$-couplings between experiment (black) and simulations (Figure S3); distributions of the dihedral angle φ of residue 3 (Figure S4); and distributions of the dihedral angle φ of residue 9 (Figure S5) (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: jochen.hub@uni-saarland.de.

ORCID

Jochen S. Hub: 0000-0001-7716-1767

Notes

The authors declare no competing financial interest.

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

■ ACKNOWLEDGMENTS

The authors thank Sarah Rauscher for sharing experimental SAXS data of the RS peptide and Po-chia Chen for helpful suggestions on the manuscript. This project was supported by the Deutsche Forschungsgemeinschaft (HU 1971/3-1 and HU 1971/4-1).

■ REFERENCES

(1) Tompa, P. Unstructural biology coming of age. *Curr. Opin. Struct. Biol.* **2011**, *21*, 419−425.
(2) Varadi, M.; Vranken, W.; Guharoy, M.; Tompa, P. Computational approaches for inferring the functions of intrinsically disordered proteins. *Front. Mol. Biosci.* **2015**, *2*, No. 45.
(3) van der Lee, R.; et al. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **2014**, *114*, 6589−6631.
(4) Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* **2014**, *83*, 553−584.
(5) Dyson, H. J.; Wright, P. E. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197−208.
(6) Ward, J.; Sodhi, J.; McGuffin, L.; Buxton, B.; Jones, D. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *J. Mol. Biol.* **2004**, *337*, 635−645.
(7) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annu. Rev. Biophys.* **2008**, *37*, 215−246.
(8) Tuukkanen, A. T.; Spilotros, A.; Svergun, D. I. Progress in small-angle scattering from biological solutions at high-brilliance synchrotrons. *IUCrJ* **2017**, *4*, 518−528.
(9) Bizien, T.; Durand, D.; Roblina, P.; Thureau, A.; Vachette, P.; Pérez, J. A Brief Survey of State-of-the-Art BioSAXS. *Protein Pept. Lett.* **2016**, *23*, 217−231.
(10) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmüller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513−5524.
(11) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmüller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2017**, *14*, 71−73.
(12) Liu, H.; Song, D.; Lu, H.; Luo, R.; Chen, H.-F. Intrinsically disordered protein-specific force field CHARMM36IDPSFF. *Chem. Biol. Drug Des.* **2018**, *92*, 1722−1735.
(13) Wang, W.; Ye, W.; Jiang, C.; Luo, R.; Chen, H.-F. New Force Field on Modeling Intrinsically Disordered Proteins. *Chem. Biol. Drug Des.* **2014**, *84*, 253−269.
(14) Song, D.; Wang, W.; Ye, W.; Ji, D.; Luo, R.; Chen, H.-F. ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins. *Chem. Biol. Drug Des.* **2017**, *89*, 5−15.
(15) Carballo-Pacheco, M.; Strodel, B. Comparison of force fields for Alzheimer's A β42: A case study for intrinsically disordered proteins. *Protein Sci.* **2017**, *26*, 174−185.
(16) Wu, H.; G. Wolynes, P.; Papoian, G. AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2018**, 11115−11125.
(17) Ye, W.; Ji, D.; Wang, W.; Luo, R.; Chen, H.-F. Test and Evaluation of ff99IDPs Force Field for Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* **2015**, *55*, 1021−1029.
(18) Zerze, G. H.; Zheng, W.; Best, R. B.; Mittal, J. Evolution of All-Atom Protein Force Fields to Improve Local and Global Properties. *J. Phys. Chem. Lett.* **2019**, *10*, 2227−2234.
(19) Olsson, S.; Frellsen, J.; Boomsma, W.; Mardia, K. V.; Hamelryck, T. Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data. *PLoS One* **2013**, *8*, No. e79439.
(20) Schröder, G. F. Hybrid methods for macromolecular structure determination: experiment with expectations. *Curr. Opin. Struct. Biol.* **2015**, *31*, 20−27.
(21) 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*, 5686−5701.
(22) Allison, J. R. Using simulation to interpret experimental data in terms of protein conformational ensembles. *Curr. Opin. Struct. Biol.* **2017**, *43*, 79−87.
(23) Sormanni, P.; et al. Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.* **2017**, *13*, 339.
(24) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106−116.
(25) Chan-Yao-Chong, M.; Durand, D.; Ha-Duong, T. Molecular Dynamics Simulations Combined with Nuclear Magnetic Resonance and/or Small-Angle X-ray Scattering Data for Characterizing Intrinsically Disordered Protein Conformational Ensembles. *J. Chem. Inf. Model.* **2019**, 1743−1758.
(26) Cesari, A.; Reißer, S.; Bussi, G. Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments. *Computation* **2018**, *6*, 15.
(27) Choy, W.-Y.; Forman-Kay, J. D. Calculation of ensembles of structures representing the unfolded state of an SH3 domain. *J. Mol. Biol.* **2001**, *308*, 1011−1032.
(28) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *J. Am. Chem. Soc.* **2007**, *129*, 5656−5664.
(29) Pelikan, M.; Hura, G.; Hammel, M. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.* **2009**, *28*, 174−189.
(30) Bertini, I.; Giachetti, A.; Luchinat, C.; Parigi, G.; Petoukhov, M. V.; Pierattelli, R.; Ravera, E.; Svergun, D. I. Conformational Space of Flexible Biological Macromolecules from Average Data. *J. Am. Chem. Soc.* **2010**, *132*, 13553−13558.
(31) Yang, S.; Blachowicz, L.; Makowski, L.; Roux, B. Multidomain assembled states of Hck tyrosine kinase in solution. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 15757−15762.
(32) Daughdrill, G. W.; Kashtanov, S.; Stancik, A.; Hill, S. E.; Helms, G.; Muschol, M.; Receveur-Bréchot, V.; Ytreberg, F. M. Understanding the structural ensembles of a highly extended disordered protein. *Mol. BioSyst.* **2011**, *8*, 308−319.
(33) Różycki, B.; Kim, Y. C.; Hummer, G. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* **2011**, *19*, 109−116.
(34) Sanchez-Martinez, M.; Crehuet, R. Application of the maximum entropy principle to determine ensembles of intrinsically disordered proteins from residual dipolar couplings. *Phys. Chem. Chem. Phys.* **2014**, *16*, 26030−26039.
(35) Tria, G.; Mertens, H. D. T.; Kachala, M.; Svergun, D. I. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2015**, *2*, 207−217.
(36) Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **2015**, *143*, No. 243150.
(37) Brookes, D. H.; Head-Gordon, T. Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **2016**, *138*, 4530−4538.
(38) Leung, H. T. A.; Bignucolo, O.; Aregger, R.; Dames, S. A.; Mazur, A.; Bernèche, S.; Grzesiek, S. A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content. *J. Chem. Theory Comput.* **2016**, *12*, 383−394.
(39) Bonvin, A. M. J. J.; Brünger, A. T. Conformational Variability of Solution Nuclear Magnetic Resonance Structures. *J. Mol. Biol.* **1995**, *250*, 80−93.

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

(40) Fennen, J.; Torda, A. E.; van Gunsteren, W. F. Structure refinement with molecular dynamics and a Boltzmann-weighted ensemble. *J. Biomol. NMR* **1995**, *6*, 163−170.
(41) Hess, B.; Scheek, R. M. Orientation restraints in molecular dynamics simulations using time and ensemble averaging. *J. Magn. Reson.* **2003**, *164*, 19−27.
(42) Best, R. B.; Vendruscolo, M. Determination of Protein Structures Consistent with NMR Order Parameters. *J. Am. Chem. Soc.* **2004**, *126*, 8090−8091.
(43) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. Mapping Long-Range Interactions in α-Synuclein using Spin-Label NMR and Ensemble Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **2005**, *127*, 476−477.
(44) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **2005**, *433*, 128−132.
(45) Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental Parameterization of an Energy Function for the Simulation of Unfolded Proteins. *Biophys. J.* **2008**, *94*, 182−192.
(46) Esteban-Martín, S.; Fenwick, R. B.; Salvatella, X. Refinement of Ensembles Describing Unstructured Proteins Using NMR Residual Dipolar Couplings. *J. Am. Chem. Soc.* **2010**, *132*, 4626−4632.
(47) Robustelli, P.; Kohlhoff, K.; Cavalli, A.; Vendruscolo, M. Using NMR Chemical Shifts as Structural Restraints in Molecular Dynamics Simulations of Proteins. *Structure* **2010**, *18*, 923−933.
(48) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **2013**, *138*, No. 094112.
(49) Beauchamp, K.; Pande, V.; Das, R. Bayesian Energy Landscape Tilting: Towards Concordant Models of Molecular Ensembles. *Biophys. J.* **2014**, *106*, 1381−1390.
(50) White, A. D.; Voth, G. A. Efficient and Minimal Method to Bias Molecular Simulations with Experimental Data. *J. Chem. Theory Comput.* **2014**, *10*, 3023−3030.
(51) White, A. D.; Dama, J. F.; Voth, G. A. Designing Free Energy Surfaces That Match Experimental Data with Metadynamics. *J. Chem. Theory Comput.* **2015**, *11*, 2451−2460.
(52) Antonov, L. D.; Olsson, S.; Boomsma, W.; Hamelryck, T. Bayesian inference of protein ensembles from SAXS data. *Phys. Chem. Chem. Phys.* **2016**, *18*, 5832−5838.
(53) Hansen, N.; Heller, F.; Schmid, N.; van Gunsteren, W. F. Time-averaged order parameter restraints in molecular dynamics simulations. *J. Biomol. NMR* **2014**, *60*, 169−187.
(54) Zhang, B.; Wolynes, P. G. Topology, structures, and energy landscapes of human chromosomes. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 6062−6067.
(55) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2016**, *2*, No. e1501177.
(56) Cesari, A.; Gil-Ley, A.; Bussi, G. Combining Simulations and Solution Experiments as a Paradigm for RNA Force Field Refinement. *J. Chem. Theory Comput.* **2016**, *12*, 6192−6200.
(57) Rangan, R.; Bonomi, M.; Heller, G. T.; Cesari, A.; Bussi, G.; Vendruscolo, M. Determination of Structural Ensembles of Proteins: Restraining vs Reweighting. *J. Chem. Theory Comput.* **2018**, 6632−6641.
(58) Best, R. B.; Zheng, W.; Mittal, J. Balanced Protein−Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* **2014**, *10*, 5113−5124.
(59) Chen, P.-c.; Hub, J. Interpretation of Solution X-Ray Scattering by Explicit-Solvent Molecular Dynamics. *Biophys. J.* **2015**, *108*, 2573−2584.
(60) Ivanović, M. T.; Bruetzel, L. K.; Lipfert, J.; Hub, J. S. Temperature-Dependent Atomic Models of Detergent Micelles Refined against Small-Angle X-Ray Scattering Data. *Angew. Chem., Int. Ed.* **2018**, *57*, 5635−5639.
(61) Hub, J. S. Interpreting solution X-ray scattering data using molecular simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 18−26.
(62) Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620−630.
(63) Kruschel, D.; Zagrovic, B. Conformational averaging in structural biology: issues, challenges and computational solutions. *Mol. BioSyst.* **2009**, *5*, 1606−1616.
(64) Eliezer, D. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23−30.
(65) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Combining Experiments and Simulations Using the Maximum Entropy Principle. *PLoS Comput. Biol.* **2014**, *10*, No. e1003406.
(66) Bottaro, S.; Bengtsen, T.; Lindorff-Larsen, K. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach. *bioRxiv* **2018**, No. 457952.
(67) Pitera, J. W.; Chodera, J. D. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445−3451.
(68) Roux, B.; Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **2013**, *138*, No. 084107.
(69) Chen, P.-c.; Hub, J. Validating Solution Ensembles from Molecular Dynamics Simulation by Wide-Angle X-ray Scattering Data. *Biophys. J.* **2014**, *107*, 435−447.
(70) Knight, C. J.; Hub, J. S. WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nucleic Acids Res.* **2015**, *43*, W225−W230.
(71) Chen, P.-c.; Hub, J. S. Structural Properties of Protein−Detergent Complexes from SAXS and MD Simulations. *J. Phys. Chem. Lett.* **2015**, *6*, 5116−5121.
(72) Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145−151.
(73) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(74) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J. Phys. Chem. B* **2015**, *119*, 5113−5123.
(75) Hess, B.; Kutzner, C.; Van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(76) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **2013**, *29*, 845−854.
(77) 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.
(78) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(79) Miyamoto, S.; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **1992**, *13*, 952−962.
(80) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1998**, *18*, 1463−1472.
(81) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(82) Van Gunsteren, W. F.; Berendsen, H. J. C. A Leap-frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1988**, *1*, 173−185.
(83) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577−2637.

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115

(84) MacKerell, A. D.; et al. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(85) Henriques, J.; Arleth, L.; Lindorff-Larsen, K.; Skepö, M. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *J. Mol. Biol.* **2018**, 2521−2539.
(86) Hess, B. Determining the shear viscosity of model liquids from molecular dynamics simulations. *J. Chem. Phys.* **2002**, *116*, 209−217.
(87) Vuister, G. W.; Bax, A. Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HNH.alpha.) coupling constants in 15N-enriched proteins. *J. Am. Chem. Soc.* **1993**, *115*, 7772−7777.

DOI: 10.1021/acs.jctc.9b00338
J. Chem. Theory Comput. 2019, 15, 5103−5115