Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations

Utsab R. Shrestha[1], Jeremy C. Smith[1,2], Loukas Petridis[1,2]*

[1]UT/ORNL Center for Molecular Biophysics, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, United States.
[2]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, United States.

*Correspondence should be addressed to: petridisl@ornl.gov

ABSTRACT

Molecular dynamics (MD) simulation is widely used to complement ensemble-averaged experiments of intrinsically disordered proteins (IDPs). However, MD often suffers from limitations of inaccuracy in the force fields and inadequate sampling. Here, we show that enhancing the sampling using Hamiltonian replica-exchange MD led to unbiased ensembles of unprecedented accuracy, reproducing small-angle scattering and NMR chemical shift experiments, for three IDPs of variable sequence properties using two recently optimized force fields. Surprisingly, we reveal that despite differences in their sequence, the inter-chain statistics of all three IDPs are similar for short contour lengths (< 10 residues).

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US DOE. The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/downloads/doe-publicaccess-plan).

INTRODUCTION

Intrinsically disordered proteins (IDPs) exhibit biological function without folding spontaneously into a unique three-dimensional (3D) structure.[1] IDPs are abundantly present in all proteomes and play major roles in signaling, transcriptional regulation and regulation of phase transitions in the cell via processes that may involve high-specificity or low-affinity interactions and recognition of partners by folding upon binding.[1-5] About 50 to 70% of the proteins in the human genome associated with cancers, diabetes, cardiovascular, and neurodegenerative diseases have a minimum of 30 residues that are intrinsically disordered, making IDPs possible drug targets.[1] Additionally, IDPs are an essential part of plant immune signaling components and also mediate plant-microbe interactions.[6]

Understanding the function of a protein requires a determination of its 3D structure.[7] IDPs adopt highly dynamic structural ensembles, which are commonly characterized by nuclear magnetic resonance (NMR)[8], small-angle X-ray/neutron scattering (SAXS/SANS),[9,10] single-molecule Förster resonance energy transfer (smFRET),[11] hydrogen-exchange mass spectrometry[12] and circular dichroism (CD).[13,14] However, the information content of the applied experimental techniques is insufficient to obtain the ensemble of 3D conformations an IDP adopts.[15] The experimental observables often represent averages over the ensemble and the data are typically sparse, providing too little information to unambiguously determine the 3D ensemble.

Molecular dynamics (MD) simulation can in principle provide the missing information and furnish a complete atomic resolution description of IDP structure and dynamics.[2] Recent optimizations of the protein and water potential energy functions[2,16-27] have led to accurate simulation of short disordered peptides and model systems.[17,18,28-31] However, the simulations are not always consistent with experiment either because of inadequate sampling or shortcomings of the force fields.[2,18,21,23,29,32,33]

A common and successful approach to determine an IDP configurational ensemble is to force the MD results to match existing experiments, either by biasing the MD potential,[34,35] or by *a posteriori* reweighting the ensemble of the MD population.[36,37] One challenge for these methods is degeneracy, *i.e.* distinct 3D conformations may yield the same observable, which may lead to over-fitting. Bayesian maximum entropy optimization approaches, which aim to perturb the MD ensemble as little as possible, have been employed to avoid over fitting.[34,37,38] However, these approaches always require a prior experimental measurement and do not afford a predictive understanding of IDPs.

Recently, by enhancing the configurational sampling of MD simulations using Hamiltonian replica-exchange MD (HREMD) the configurational ensemble of an IDP was generated that is in quantitative agreement to SAXS, SANS and NMR measurements without biasing or reweighting the simulations.[39,40] HREMD improves sampling by scaling the intra-protein and protein-water potentials[16,19] of higher-order replicas, while keeping the potential of the lowest rank replica unscaled[41-44] so as to represent the physically-meaningful interactions of the system. However, two IDPs[39,40] were studied and the general applicability of this approach has not been established.

Here, we report that HREMD produces configurational ensembles consistent with SAXS, SANS and NMR experiments for three IDPs with markedly different sequence characteristics: Histatin 5 (24 residues) and Sic 1 (92 residues), both of which have an abundance of positively charged residues, and the N-terminal SH4UD (95 residues) of c-Src kinase, which contains positively and negatively charged residues mixed over the sequence. The HREMD results are in agreement with experimental data on both local and global properties when employing either of two force fields (Amber ff03ws[19] with TIP4P/2005s[19] and Amber ff99SB-*disp*[16] with modified TIP4P-D,[16] hereafter termed as a03ws and a99SB-disp respectively). In contrast, standard MD simulations of equivalent computational cost as HREMD generate ensembles consistent only with NMR, but not with SAXS. Further, the HREMD ensembles of IDPs are found to be described by a theoretical semiflexible polymer chain model quantifying the stiffness and strength of interaction with solvent. We suggest "best practices" in achieving accurate and efficient IDP sampling using HREMD and discuss differences in the size between Sic 1 and SH4UD. The results demonstrate quite clearly that the recently optimized force fields are reliable and that the current major impediment to accurate simulation of IDPs using is sampling. HREMD is therefore the present tool of choice for obtaining atomic-detailed IDP ensembles.

RESULTS

HREMD ensembles in agreement with SAXS, SANS and NMR.

We conducted HREMD simulations of three IDPs with varying amino acid composition (**Fig. S1**), employing two force fields: a03ws[19] and a99SB-disp[16]. For comparison, we also conducted standard MD, *i.e.* without enhancing the sampling, of the same cumulative length as the HREMD (**Tables S1-S4**). The histograms of a radius of gyration (R_g) show the IDPs adopt a continuum of collapsed to extended structures (**Fig. 1a-c**).

The global, ensemble-averaged properties of IDPs such as R_g, shape, and chain statistics can be derived using small-angle scattering. We calculated the ensemble-averaged theoretical SAXS and SANS curves from the simulation trajectories, by taking into account explicitly the protein hydration shell and without reweighting, and compared them directly to the experiments.

We found an excellent agreement of the HREMD-generated ensembles with SAXS and SANS measurements for both force fields (SAXS in **Fig. 1d-f** and SANS in **Fig. S2**), whereas the standard MD simulations were found to deviate from the experiments, except for Sic 1 with a03ws. The agreement between simulation and experiment was quantified with the χ^2 value as defined in **Eq. (5)** and listed in **Table S5**. The histograms of R_g show that standard MD simulations sample more compact structures than does HREMD with the same force fields. Therefore, for the IDPs studied here, poor agreement with experiment arises primarily from insufficient sampling rather than from shortcomings of the force fields.

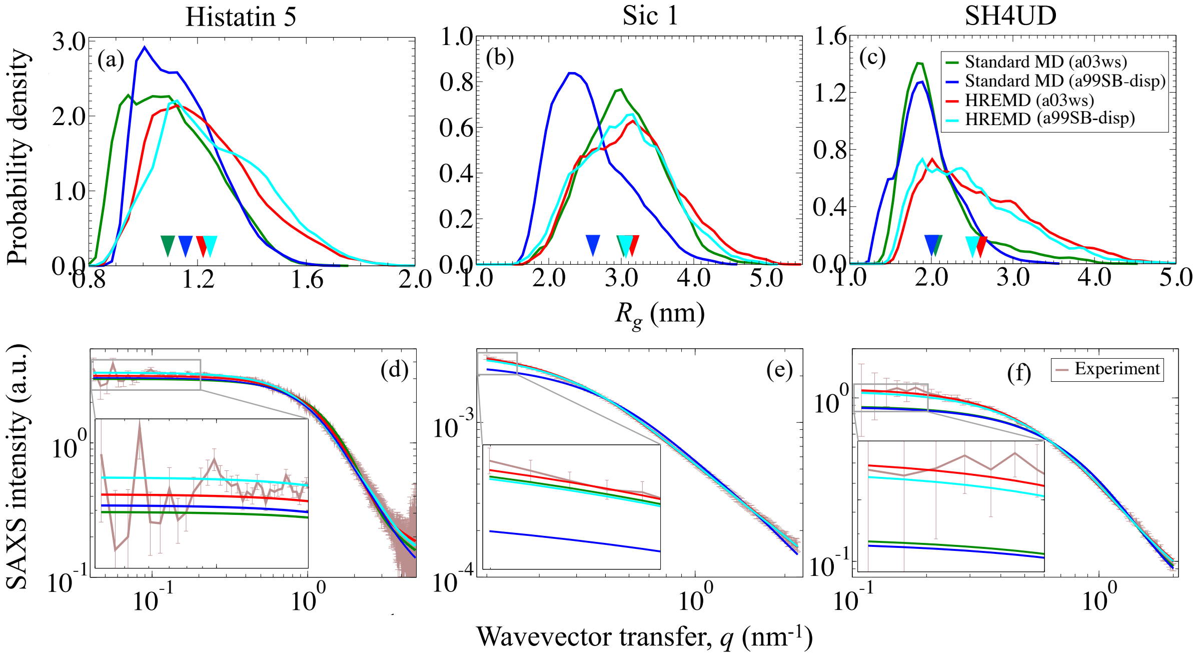

Fig. 1. (a-c) The histograms of R_g of (a) Histatin 5, (b) Sic 1 and (c) SH4UD obtained from MD simulations. The inverted triangles indicate the average R_g of each simulation. (d-f) The SAXS profiles calculated from simulations (using SWAXS[45]) are compared to experiments for (d) Histatin 5,[30] (e) Sic 1,[46] and (f) SH4UD.[40] Insets: SAXS data are zoomed at low-q values to show the differences in intensity for different force fields and sampling methods. In all cases the color code indicates the force fields, a03ws[19] or a99SB-disp,[16] and sampling methods, standard MD or HREMD (**Tables S1 and S2**). HREMD results are from the lowest rank replica of the simulations shown by bold-italics font in **Table S2**. SANS data of SH4UD are shown in Fig. S2.

NMR chemical shifts provide information on the local chemical environment of protein atoms and reflect structural factors such as backbone and side-chain conformations. To further validate the simulations, we calculated the ensemble-averaged backbone chemical shifts (C^{α}, C^{β} and N^{H} for Sic 1 and SH4UD, and H^{N} and H^{α} for Histatin 5) and compared to previously reported experiments (**Figs. 2, S3-S6**). The agreement between theoretical and experimental NMR chemical shifts was quantified by calculating the mean normalized deviation as defined by **Eq. (6)**. For Sic 1 and SH4UD, we found an excellent agreement with the experiments for all force fields and sampling methods, whereas the agreement is not quite as good for Histatin 5 (**Figs. 2 and S3-S6**).

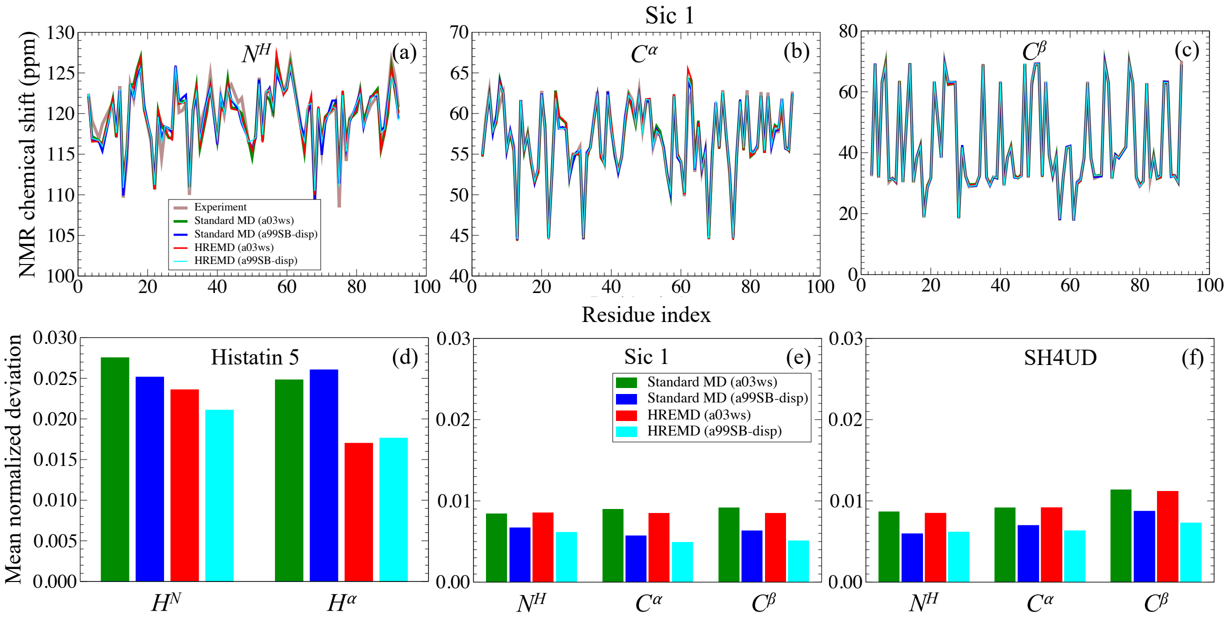

Fig. 2. Comparison between the ensemble-averaged experimental and calculated NMR chemical shifts of backbone atoms (a) N^{H}, (b) C^{α} and (c) C^{β}, for Sic 1. The mean normalized deviation of MD-derived NMR chemical shifts of backbone atoms with respect to experimental values, as defined in **Eq. (6)**, for (d) Histatin 5,[47] (e) Sic 1,[46] and (f) SH4UD.[48] The color code indicates the force field and sampling method used. The theoretical NMR chemical shifts are calculated using SHIFTX2,[49] which has relatively high values of root mean square errors of 0.1711 ppm and 0.1231 ppm for H^{N} and H^{α} respectively compared to N^{H}, C^{α} and C^{β}.

Both force fields and sampling methods predict nearly the same transient secondary structure elements. Transient helices, which are considered to be biologically relevant,[50-52] were found proximal to known phosphorylation residues of Sic 1[46] and to known lipid-binding or phosphorylation residues in SH4UD.[48,53] In contrast, the propensity of each secondary structure element is found to depend on both the force fields and sampling methods (**Figs. S7 and S8**). The IDPs we studied mostly showed a high propensity for coils that lack secondary structure, consistent with the lack of long-range contacts found in the simulations (**Fig. S9**).

Polymer properties.

We estimated the stiffness of the protein backbones by calculating the orientational correlation function

$$C(s) = < n_i . n_{i+s} > \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where s=|i-j| is the pairwise residue separation (sometimes called contour length), and n_i is the unit vector connecting the backbone atoms N and C of residue *i*. The steeper the decay of *C(s)*, the lower the stiffness of the chain. *C(s)* is similar for the three IDPs for $s\leq10$, exhibiting an exponential decay $C(s) = e^{-s/l_p}$, where l_p is the persistence length. l_p provides the maximum size of a protein segment over which the structural fluctuations are correlated. In other words, it is the measure of stiffness of a polypeptide chain. We found l_p ~ 1 nm for all IDPs, in good agreement to the values for intrinsically disordered proteins.[54,55] A power law decay ($\sim s^{-3/2}$) is found for Sic 1 at $10<s\leq26$, whereas correlations decay more rapidly and vanish for $s>10$ for Histatin 5 and SH4UD. Therefore, Sic 1 is the stiffest.

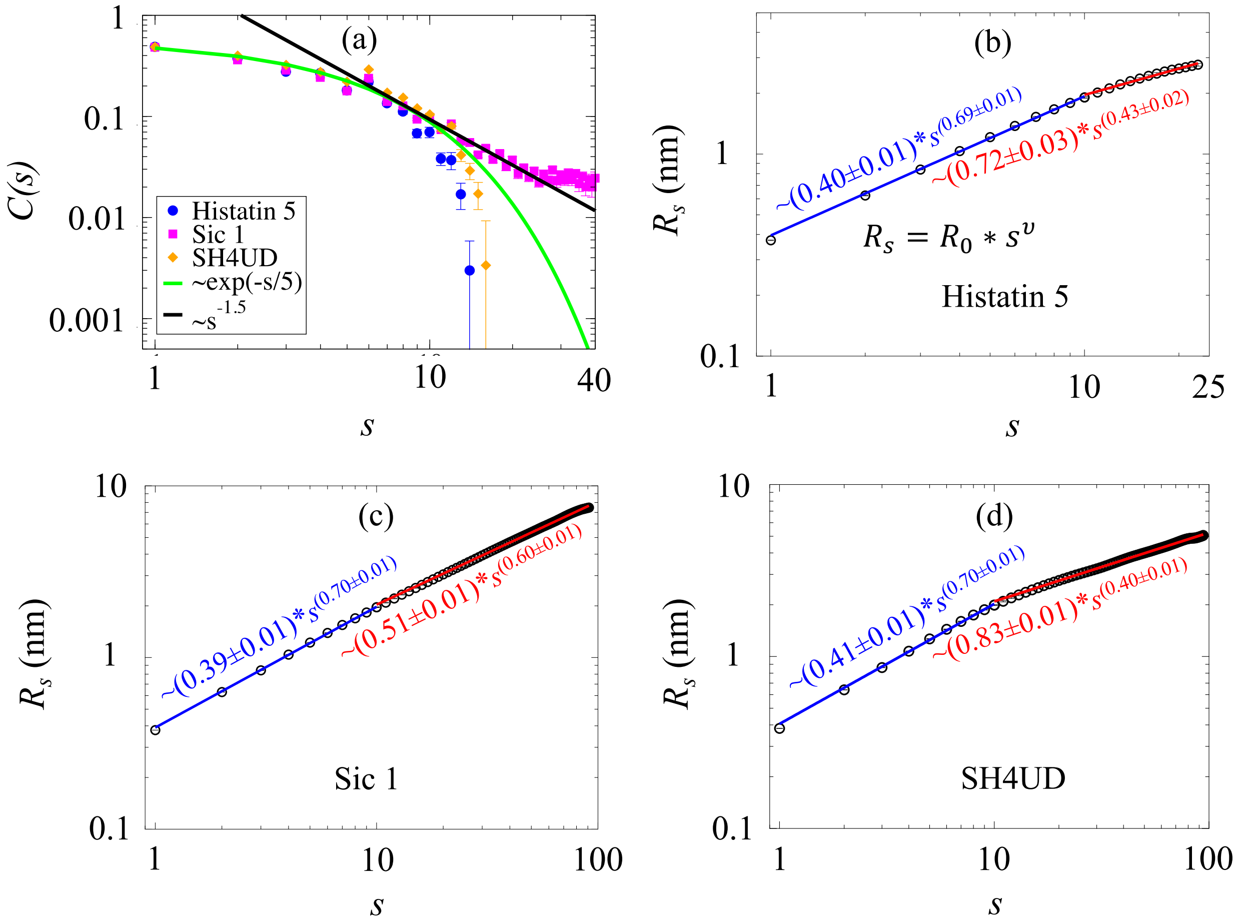

Fig. 3. Chain statistics of IDPs. (a) The orientational correlation function as a function of the pairwise residue sequence separation, *s*. For $s \leq 10$, $C(s)$ is fitted by $C(s) = e^{-s/l_p}$ for each IDP, which estimates the persistence length (l_p). For *s*>10 the power law $C(s) \sim s^{-3/2}$ applies only for Sic 1, whereas for Histatin 5 and SH4UD the correlation vanishes. (b-d) The average pairwise geometric distance (R_s) between C-alpha atoms of two residues at separation *s* for (b) Histatin 5, (c) Sic 1 and (d) SH4UD. The data are fitted by Eq. (2) in two regimes, *s*≤10 (blue) and *s*>10 (red). The error bars are smaller than the symbol size.

The statistics of internal distances ("scaling properties") of polymers in dilute solution can be characterized using the Flory scaling law given by Eq. (2):

$$R_s = R_0\, s^{\upsilon} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

where R_s is the average intraprotein pairwise distance between the C$_\alpha$ atoms of residues *i* and *j* at pairwise separation *s*=|*i*-*j*|, the prefactor R_0 is a constant and ν is the Flory exponent. Balanced polymer-solvent and intrapolymer interactions give rise to Gaussian coil and ν=0.5, while a self-avoiding random walk (SARW) with ν=0.588 is predicted when the polymer-water interactions are favored. Interestingly, we found two different power law regimes are needed to fit the data[56,57] (**Fig. 3b-d**). At short contour lengths (s≤10), R_s is similar for all three IDPs, with $\nu \approx 0.70$, which indicates chain configurations stiffer than a SARW, and with a prefactor of R_0 ~0.4 nm (R_0 is the average distance between two consecutive C-alpha atoms). On the other hand, at longer residue separations (s>10) the R_s of the three IDPs deviate. Histatin 5 and SH4UD with $\nu \approx 0.43$ and 0.40, respectively, adopt more collapsed global conformations than SARW. In contrast, Sic 1 ($\nu \approx 0.60$) remains stiff even at longer residue separations.

DISCUSSION

IDPs present a new paradigm for understanding flexibility-function relationships in biology.[1,58-60] Currently, it is not possible to determine the ensemble of the 3D structures that an IDP adopts from either experiment or simulation alone. The number of experimental observables is considerably smaller than the number of the IDP's configurational degrees of freedom, making model reconstruction from experimental data a highly underdetermined problem. For MD simulations, although improved molecular mechanics methods perform well for small model disordered peptides[2,18,28,30,31], it has often been necessary to bias or reweight the MD results to achieve consistency with experiments.[34,35,37,38,61-63] The reason MD has not always been accurate is unclear: it could be deficiencies in the force fields, insufficient sampling, or both.

Here, we demonstrated that HREMD reproduces key experimental observables (SAXS, SANS and NMR) using two different force fields for three different IDPs. In contrast, the ensemble generated by standard MD of equivalent length failed to match SAXS data (**Fig. 1**).

The comparison of standard MD and HREMD using the same force field suggests the a03ws and a99SB-disp force fields are of adequate accuracy and that enhanced sampling techniques are necessary to reproduce experimental data.

We found that the calculated NMR chemical shifts and the *loci* of secondary structure elements are the easiest to converge as they are consistent between all the simulations, independent of force field and sampling method. In contrast, HREMD is required for SAXS observables to converge to the experimental values. The most difficult quantities to converge are the secondary structure propensities, which were found here to depend on both the force field and the sampling method, perhaps more on the former than the latter **(Fig. S7 and S8),** with a03ws and a99SB-disp having biases towards helices and β-sheets, respectively.

The data show that MD simulations can be in *apparent* agreement with NMR chemical shifts, which measure local structural information,[64] while failing to reproduce SAXS/SANS intensities, which determine with high precision more global structural properties (here distributions of distances between pairs of nuclei that are more than ~1 nm apart[61,65]) (**Figs. 1, 2 and S10**).[40] Thus, agreement with NMR alone is not always a definitive test of the accuracy of MD simulations of IDPs. It is critical to analyze and compare both local and global properties[16,66] of IDPs to ensure the simulations have indeed generated accurate ensembles.

Simple theories established for semiflexible homopolymers and heteropolymers have been shown to provide a qualitative description of IDP structural properties such as stiffness[67-69] and solvent quality.[11,13,70-72] The high fidelity HREMD trajectories reveal that, despite having markedly different sequences, the IDPs studied here have a common hierarchical chain architecture. For short contour lengths (up to ~10 residues) the chain statistics of all three IDPS are similar, as evidenced by R_s and *C(s)*. These short segments are relatively stiff with a Flory exponent of ν~0.7. Beyond this critical contour length, the IDPs differentiate. SH4UD and Histatin 5 become flexible, while Sic 1 remains relatively stiff with power-law decay in *C(s)* that implies long-range spatial correlations.[68] This is consistent with Sic 1 being more extended than SH4UD.

The origin of the stiffness of Sic 1 relative to SH4UD can be understood by examining their primary sequences (**Fig. S1**). All the charged residues of Sic 1 are positive, leading to electrostatic repulsion between them. Further, Sic 1 contains 15 proline and 5 glycine residues. Proline is stiff due to its cyclic sidechain, whereas the absence of a sidechain for glycine increases backbone flexibility, which is known to be disorder-promoting.[55,73] In comparison, SH4UD has both positively and negatively charged residues, 11 prolines and 12 glycines.

We now discuss the HREMD method[41,42,44] and make recommendations for its optimal use in IDPs. HREMD enhances sampling by changing the quality of water as a good solvent for an IDP. This is achieved by effectively heating up only the solute by scaling the intraprotein and protein-solvent potential energy functions. An exchange of coordinates is allowed between neighboring replicas if the Monte Carlo metropolis criterion is satisfied.[41,42] The HREMD method was chosen because it does not necessitate a predefined reaction coordinate. The advantage of HREMD over temperature replica exchange MD is that HREMD crosses entropic barriers[74] more efficiently and a smaller number of replicas is sufficient, *i.e.* is computationally more efficient.

The total number of replicas (n) used, the scaling factor (λ_i) or the effective temperature (T_i) of a replica and the average exchange probability (p_{ex}) of the lowest rank replica are listed in **Tables S2-S4**. A T_{max} of 400-450 K (lower limit) being needed, similar to T_{max} = 400 K used in previous studies,[39,75,76] and p_{ex} ranging from 0.3 to 0.5. Moreover, to estimate the upper limit of effective temperature, we performed HREMD of Histatin 5 using a99SB-disp, T_{max} = 800 K and 24 replicas (**Table S3**). This simulation generated the ensemble in the lowest rank replica similar to that of HREMD with T_{max} = 450 K (**Fig. S11a**). However, we noted that replica from T_i = 522 K and above sampled collapsed structures when compared to the ensemble of the lowest rank replica. Therefore, we suggest 450 K<T_{max}<500 K is an appropriate choice for the upper limit of maximum effective temperature (**Fig. S11a**). However, choosing the higher value of T_{max} would increase the number of replicas and thus computational cost.

In summary, we demonstrate HREMD simulations as an effective method to generate accurate structural ensembles of three IDPs with varying amino acid composition (Histatin 5, Sic 1 and SH4UD). The unbiased HREMD trajectories, calculated without using any experimental input or predefined reaction coordinate, are in excellent agreement with SAXS, SANS and NMR observables. Nonetheless, comparison to experimental data was imperative to confirm the accuracy of MD results. Moreover, HREMD simulations performed using two recent molecular mechanics force fields (a03ws and a99SB-disp) converge to the same distribution of R_g. In contrast, neither of the force fields could reproduce SAXS experiments with standard MD of the same cumulative length as HREMD. The results suggest adequately sampled simulations using recent IDP specific force fields can reliably generate the 3D ensembles of IDPs (**Fig. S12**), which is a prerequisite to an understanding of the biological function of IDPs. We also report that despite differences in their sequence, all three IDPs have similar local chain statistics for short lengths (less than ~ 10 residues). More studies are required to establish whether this is a universal IDP behavior.

MATERIALS AND METHODS

Experimental SAXS and NMR data.

The experimental SAXS data of Histatin 5, Sic 1 and SH4UD were taken from Henriques et. al. (2015),[30] Protein Ensemble Database (http://pedb.vib.be)[46] and our previous work[40] respectively. Similarly, NMR chemical shifts of backbone atoms, (C^{α}, C^{β}, N^{H}, H^{α}, H^{N}) of Histatin 5, Sic 1 and SH4UD were acquired from the literature,[47] Protein Ensemble Database[46] and Biological Magnetic Resonance Data Bank (BMRB) database entry 15563[48] respectively.

MD simulations.

The initial 3D structures of IDPs were obtained from I-TASSER.[77] An MD-equilibrated starting structure with R_g value close to experimental SAXS was chosen for the production simulation of each IDP. The same starting structure of IDP was utilized for each force field and sampling method.

We performed standard molecular dynamics simulations with two recently optimized force fields, Amber ff03ws[19,78,79] with TIP4P/2005s[19] (a03ws) and Amber ff99SB-*disp*[16,80] with the modified TIP4P-D[16,21] water model (a99SB-disp) using GROMACS.[81-86] All bonds involving hydrogen atoms were constrained using LINCS algorithm.[87] The Verlet leapfrog algorithm was used to numerically integrate the equation of motions with a time step of 2 fs. A cutoff of 1.2 nm was used for short-range electrostatic and Lennard-Jones interactions. Long-range electrostatic interactions were calculated by particle-mesh Ewald[88] summation with a fourth order interpolation and a grid spacing of 0.16 nm. The solute and solvent were coupled separately to a temperature bath of 300 K, 293 K and 300 K for Histatin 5, Sic 1 and SH4UD respectively to match the temperatures measured at the experiments using modified Berendsen thermostat with a relaxation time of 0.1 ps. The pressure coupling was fixed at 1 bar using Parrinello-Rahman algorithm[89] with a relaxation time of 2 ps and isothermal compressibility of 4.5*10^{-5} bar^{-1}. The energy of each system was minimized using 1000 steepest decent steps followed by 1 ns equilibration at NVT and NPT ensembles. The production runs were carried out in the NPT ensemble.

Enhanced sampling MD simulations.

We employed replica-exchange with solute tempering 2 (REST2),[41,42] a Hamiltonian Replica-Exchange MD (HREMD) simulation method to enhance the conformational sampling. REST2 is implemented in GROMACS[81-86] patched with PLUMED.[90] The interaction potentials of intraprotein and protein-solvent were scaled by a factor λ and $\sqrt{\lambda}$ respectively, while water-water interactions were unaltered.[41,42,76,91] The scaling factor λ_i, and corresponding effective temperatures T_i of the ith replica are given by,

$$\lambda_i = \frac{T_0}{T_i} = \exp\left(-\frac{i}{(n-1)}\ln\left(\frac{T_{max}}{T_0}\right)\right)\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(3)$$

where T_0 and T_{max} are the effective temperatures of lowest rank (unscaled) and the highest rank replicas respectively, and n is the total number of replicas used. For analysis we use only the trajectory of the unscaled for lowest rank replica (λ_0=1 or T_0). Exchange of coordinate between neighboring replicas was attempted every 400 MD steps. The details of HREMD and standard MD simulations are shown in **Tables S1-S4**. The secondary structure prediction was calculated with DSSP.[92]

Error analysis.

To estimate the error from HREMD trajectory, we divided the trajectory into five equal blocks each containing 10,000 frames (0-100, 100-200, 200-300, 300-400 and 400-500 ns). The mean value for each block, m_i (i=1 to 5), was first calculated. The reported error bars are the standard error of the mean of the (m_1, m_2, m_3, m_4, m_5) distribution.

$$\text{i.e., Error bar } = \sqrt{\frac{1}{n(n-1)}\sum_{i}^{n=5}(m_i - \overline{m})^2} \qquad \text{……………………………………..(4)}$$

where $\overline{m}$ is the mean value and n=5 is the number of blocks used.

Theoretical SAXS profiles.

The theoretical SAXS and SANS intensities were calculated with SWAXS[45,93] and SASSENA,[94] respectively, by taking into account of explicit hydration water, which contributes to the signal.[45] The agreement between experiment and simulation was determined by a χ^2 value:

$$\chi^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left\{\frac{[<I_{expt}(q_i)>-(c<I_{sim}(q_i)>+bgd)]}{\sigma_{expt}(q_i)}\right\}^2 \text{ ………………………………… (5)}$$

where $<I_{expt}(q)>$ and $<I_{sim}(q)>$ are the ensemble averaged experimental and theoretical SAXS data, respectively, k is the number of experimental q points, c is a scaling factor, bgd is a constant background and σ_{expt} is the experimental error. In Eq. (4), c is a factor to scale calculated values to the experiment because the experimental values are often expressed in arbitrary units. It does not change the shape of the SAXS curve. Similarly, bgd is used to incorporate the uncertainty due to mismatch in buffer subtraction at higher q-values[13] in the experiment.

Theoretical NMR chemical shifts.

Theoretical NMR chemical shifts were calculated with SHIFTX2[49] by taking the average over all frames from the MD trajectory. The discrepancy between calculated and experimental values are measured by Mean Normalized Deviation is defined as,

$$Mean\ Normalized\ Deviation = \frac{1}{n}\sum_{i=1}^{n}\frac{|CS_i^{expt} - (CS_i^{calc} - offset)|}{CS_i^{expt}} \quad \dots\dots (6)$$

where CS_i^{calc} and CS_i^{expt} are the theoretical and experimental NMR chemical shift values respectively of residue index *i* of a protein with *n* number of residues, *offset* obtained from linear regression analysis are used for each backbone atom (N^H, C^α, C^β) and IDP (Histatin 5, Sic 1, SH4UD) as shown in **Figs. S3-S5** and |…| is the modulus of the value enclosed.

CONFLICT OF INTEREST

Authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank Dr. Marie Skepö for generously sharing the SAXS data of Histatin 5. This research was supported by project ERKP300 funded by the Office of Biological & Environmental Research in the Department of Energy Office of Science. This research used a computing resource of the Oak Ridge Leadership Computing Facility (ALCC) and of the CADES at the Oak Ridge National Laboratory, which is supported by the Office of Science of DOE under Contract No. DEAC05-00OR22725.

REFERENCES

(1) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *Annual Review of Biophysics* **2008**, *37*, 215-246.

(2) Chong, S.-H.; Chatterjee, P.; Ham, S. *Annu. Rev. Phys. Chem.* **2017**, *68*, 117-134.

(3) Sun, X.; Rikkerink, E. H.; Jones, W. T.; Uversky, V. N. *Plant Cell* **2013**, *25*, 38-55.

(4) Mitrea, D. M.; Cika, J. A.; Guy, C. S.; Ban, D.; Banerjee, P. R.; Stanley, C. B.; Nourse, A.; Deniz, A. A.; Kriwacki, R. W. *eLife* **2016**, *5*, e13571.

(5) Martin, E. W.; Mittag, T. *Biochemistry* **2018**, *57*, 2478-2487.

(6) Marín, M.; Ott, T. *Chem. Rev.* **2014**, *114*, 6912-6932.

(7) van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R. J.; Daughdrill, G. W.; Dunker, A. K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D. T.; Kim, P. M.; Kriwacki, R. W.; Oldfield, C. J.; Pappu, R. V.; Tompa, P.; Uversky, V. N.; Wright, P. E.; Babu, M. M. *Chem. Rev.* **2014**, *114*, 6589-6631.

(8) Dyson, H. J.; Wright, P. E. *J. Biomol. NMR* **2019**, *73*, 651-659.

(9) Cordeiro, T. N.; Herranz-Trillo, F.; Urbanek, A.; Estaña, A.; Cortés, J.; Sibille, N.; Bernadó, P. *Curr. Opin. Struct. Biol.* **2017**, *42*, 15-23.

(10) Mansouri, A. L.; Grese, L. N.; Rowe, E. L.; Pino, J. C.; Chennubhotla, S. C.; Ramanathan, A.; O'Neill, H. M.; Berthelier, V.; Stanley, C. B. *Molecular BioSystems* **2016**, *12*, 3695-3701.

(11) Schuler, B.; Soranno, A.; Hofmann, H.; Nettels, D. *Annu Rev Biophys* **2016**, *45*, 207-231.

(12) Balasubramaniam, D.; Komives, E. A. *Biochim. Biophys. Acta* **2013**, *1834*, 1202-1209.

(13) Riback, J. A.; Bowman, M. A.; Zmyslowski, A. M.; Knoverek, C. R.; Jumper, J. M.; Hinshaw, J. R.; Kaye, E. B.; Freed, K. F.; Clark, P. L.; Sosnick, T. R. *Science* **2017**, *358*, 238.

(14) Na, J. H.; Lee, W. K.; Yu, Y. G. *Int J Mol Sci* **2018**, *19*.

(15) Baker, C. M.; Best, R. B. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2014**, *4*, 182-198.

(16) Robustelli, P.; Piana, S.; Shaw, D. E. *Proc Natl Acad Sci U S A* **2018**, *115*, E4758-E4766.

(17) Huang, J.; MacKerell, A. D., Jr. *Curr. Opin. Struct. Biol.* **2017**, *48*, 40-48.

(18) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D., Jr. *Nat. Methods* **2017**, *14*, 71-73.

(19) Best, R. B.; Zheng, W.; Mittal, J. *J Chem Theory Comput* **2014**, *10*, 5113-5124.

(20) Lindorff-Larsen, K.; Trbovic, N.; Maragakis, P.; Piana, S.; Shaw, D. E. *J. Am. Chem. Soc.* **2012**, *134*, 3787-3791.

(21) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. *J. Phys. Chem. B* **2015**, *119*, 5113-5123.

(22) Shabane, P. S.; Izadi, S.; Onufriev, A. V. *J Chem Theory Comput* **2019**.

(23) Chan-Yao-Chong, M.; Durand, D.; Ha-Duong, T. *J Chem Inf Model* **2019**, *59*, 1743-1758.

(24) Yu, L.; Li, D. W.; Bruschweiler, R. *J Chem Theory Comput* **2020**.

(25) Best, R. B. *Curr. Opin. Struct. Biol.* **2017**, *42*, 147-154.

(26) Best, R. B. *Curr. Opin. Struct. Biol.* **2019**, *60*, 27-38.

(27) Zerze, G. H.; Zheng, W.; Best, R. B.; Mittal, J. *J Phys Chem Lett* **2019**, *10*, 2227-2234.

(28) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmuller, H. *J Chem Theory Comput* **2015**, *11*, 5513-5524.

(29) Henriques, J.; Arleth, L.; Lindorff-Larsen, K.; Skepo, M. *J. Mol. Biol.* **2018**, *430*, 2521-2539.

(30) Henriques, J.; Cragnell, C.; Skepo, M. *J Chem Theory Comput* **2015**, *11*, 3420-3431.

(31) Henriques, J.; Skepo, M. *J Chem Theory Comput* **2016**, *12*, 3407-3415.

(32) Lincoff, J.; Sasmal, S.; Head-Gordon, T. *J. Chem. Phys.* **2019**, *150*, 104108.

(33) Bhowmick, A.; Brookes, D. H.; Yost, S. R.; Dyson, H. J.; Forman-Kay, J. D.; Gunter, D.; Head-Gordon, M.; Hura, G. L.; Pande, V. S.; Wemmer, D. E.; Wright, P. E.; Head-Gordon, T. *J. Am. Chem. Soc.* **2016**, *138*, 9730-9742.

(34) Hermann, M. R.; Hub, J. S. *J Chem Theory Comput* **2019**, *15*, 5103-5115.

(35) Roux, B.; Weare, J. *J. Chem. Phys.* **2013**, *138*, 084107.

(36) Fisher, C. K.; Huang, A.; Stultz, C. M. *J. Am. Chem. Soc.* **2010**, *132*, 14919-14927.

(37) Crehuet, R.; Buigues, P. J.; Salvatella, X.; Lindorff-Larsen, K. *Entropy* **2019**, *21*.

(38) Hummer, G.; Kofinger, J. *J. Chem. Phys.* **2015**, *143*, 243150.

(39) Liu, X.; Chen, J. *J Chem Theory Comput* **2019**, *15*, 4708-4720.

(40) Shrestha, U. R.; Juneja, P.; Zhang, Q.; Gurumoorthy, V.; Borreguero, J. M.; Urban, V.; Cheng, X.; Pingali, S. V.; Smith, J. C.; O'Neill, H. M.; Petridis, L. *Proc Natl Acad Sci U S A* **2019**.

(41) Wang, L.; Friesner, R. A.; Berne, B. J. *J. Phys. Chem. B* **2011**, *115*, 9431-9438.

(42) Bussi, G. *Mol. Phys.* **2013**, *112*, 379-384.

(43) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141-151.

(44) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proc Natl Acad Sci U S A* **2005**, *102*, 13749-13754.

(45) Chen, P. C.; Hub, J. S. *Biophys. J.* **2014**, *107*, 435-447.

(46) Mittag, T.; Marsh, J.; Grishaev, A.; Orlicky, S.; Lin, H.; Sicheri, F.; Tyers, M.; Forman-Kay, J. D. *Structure* **2010**, *18*, 494-506.

(47) Brewer, D.; Hunter, H.; Lajoie, G. *Biochem. Cell Biol.* **1998**, *76*, 247-256.

(48) Pérez, Y.; Gairí, M.; Pons, M.; Bernadó, P. *J. Mol. Biol.* **2009**, *391*, 136-148.

(49) Han, B.; Liu, Y.; Ginzinger, S. W.; Wishart, D. S. *J. Biomol. NMR* **2011**, *50*, 43-57.

(50) Kennedy, J. A.; Daughdrill, G. W.; Schmidt, K. H. *Nucleic Acids Res* **2013**, *41*, 10215-10227.

(51) Wright, P. E.; Dyson, H. J. *Nature reviews. Molecular cell biology* **2015**, *16*, 18-29.

(52) Hendus-Altenburger, R.; Lambrughi, M.; Terkelsen, T.; Pedersen, S. F.; Papaleo, E.; Lindorff-Larsen, K.; Kragelund, B. B. *Cell. Signal.* **2017**, *37*, 40-51.

(53) Pérez, Y.; Maffei, M.; Igea, A.; Amata, I.; Gairi, M.; Nebreda, A. R.; Bernado, P.; Pons, M. *Sci Rep* **2013**, *3*, 1295.

(54) Chin, A. F.; Toptygin, D.; Elam, W. A.; Schrank, T. P.; Hilser, V. J. *Biophys. J.* **2016**, *110*, 362-371.

(55) Cheng, S.; Cetinkaya, M.; Grater, F. *Biophys. J.* **2010**, *99*, 3863-3869.

(56) Gomes, G. N.; Krzeminski, M.; Martin, E. W.; Mittag, T.; Head-Gordon, T.; Forman-Kay, J. D.; Gradinaru, C. C. **2020**.

(57) Valleau, J. P. *The Journal of Chemical Physics* **1996**, *104*, 3071-3074.

(58) Knowles, T. P. J.; Vendruscolo, M.; Dobson, C. M. *Nature Reviews Molecular Cell Biology* **2014**, *15*, 384.

(59) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc Natl Acad Sci U S A* **2008**, *105*, 5762-5767.

(60) Uversky, V. N.; Roman, A.; Oldfield, C. J.; Dunker, A. K. *Journal of Proteome Research* **2006**, *5*, 1829-1842.

(61) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. *Journal of the American Chemical Society* **2007**, *129*, 5656-5664.

(62) Cavalli, A.; Camilloni, C.; Vendruscolo, M. *The Journal of Chemical Physics* **2013**, *138*, 094112.

(63) Pitera, J. W.; Chodera, J. D. *Journal of Chemical Theory and Computation* **2012**, *8*, 3445-3451.

(64) Jensen, M. R.; Zweckstetter, M.; Huang, J. R.; Blackledge, M. *Chem. Rev.* **2014**, *114*, 6632-6660.

(65) Neylon, C. *Eur. Biophys. J.* **2008**, *37*, 531-541.

(66) Demerdash, O.; Shrestha, U. R.; Petridis, L.; Smith, J. C.; Mitchell, J. C.; Ramanathan, A. *Front Mol Biosci* **2019**, *6*, 64.

(67) Shrestha, U. R.; Smith, S.; Pingali, S. V.; Yang, H.; Zahran, M.; Breunig, L.; Wilson, L. A.; Kowali, M.; Kubicki, J. D.; Cosgrove, D. J.; O'Neill, H. M.; Petridis, L. *Cellulose* **2019**, *26*, 2267-2278.

(68) Baschnagel, J.; Meyer, H.; Wittmer, J.; Kulic, I.; Mohrbach, H.; Ziebert, F.; Nam, G. M.; Lee, N. K.; Johner, A. *Polymers (Basel)* **2016**, *8*.

(69) Ghosh, A.; Gov, N. S. *Biophys. J.* **2014**, *107*, 1065-1073.

(70) Fuertes, G.; Banterle, N.; Ruff, K. M.; Chowdhury, A.; Mercadante, D.; Koehler, C.; Kachala, M.; Estrada Girona, G.; Milles, S.; Mishra, A.; Onck, P. R.; Grater, F.; Esteban-Martin, S.; Pappu, R. V.; Svergun, D. I.; Lemke, E. A. *Proc Natl Acad Sci U S A* **2017**, *114*, E6342-E6351.

(71) Hofmann, H.; Soranno, A.; Borgia, A.; Gast, K.; Nettels, D.; Schuler, B. *Proc Natl Acad Sci U S A* **2012**, *109*, 16155-16160.

(72) Das, R. K.; Pappu, R. V. *Proceedings of the National Academy of Sciences* **2013**, *110*, 13392-13397.

(73) Rauscher, S.; Baud, S.; Miao, M.; Keeley, F. W.; Pomes, R. *Structure* **2006**, *14*, 1667-1676.

(74) Nymeyer, H. *Journal of Chemical Theory and Computation* **2008**, *4*, 626-636.

(75) Graen, T.; Klement, R.; Grupi, A.; Haas, E.; Grubmuller, H. *Chemphyschem* **2018**, *19*, 2507-2511.

(76) Peng, E.; Todorova, N.; Yarovsky, I. *PLoS One* **2017**, *12*, e0186219.

(77) Roy, A.; Kucukural, A.; Zhang, Y. *Nat Protoc* **2010**, *5*, 725-738.

(78) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999-2012.

(79) Best, R. B.; Mittal, J. *The Journal of Physical Chemistry B* **2010**, *114*, 14916-14923.

(80) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Structure, Function, and Bioinformatics* **2006**, *65*, 712-725.

(81) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43-56.

(82) Lindahl, E.; Hess, B.; van der Spoel, D. *Molecular modeling annual* **2001**, *7*, 306-317.

(83) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701-1718.

(84) Suardíaz, R.; Pérez, C.; Crespo-Otero, R.; García de la Vega, J. M.; Fabián, J. S. *Journal of Chemical Theory and Computation* **2008**, *4*, 448-456.

(85) 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. *Bioinformatics* **2013**, *29*, 845-854.

(86) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. *SoftwareX* **2015**, *1–2*, 19-25.

(87) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. *J. Comput. Chem.* **1997**, *18*, 1463-1472.

(88) Darden, T.; York, D.; Pedersen, L. *The Journal of Chemical Physics* **1993**, *98*, 10089-10092.

(89) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182-7190.

(90) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. *Comput. Phys. Commun.* **2009**, *180*, 1961-1972.

(91) Sugita, Y.; Okamoto, Y. *Chemical physics letters* **1999**, *314*, 141-151.

(92) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577-2637.

(93) Park, S.; Bardhan, J. P.; Roux, B.; Makowski, L. *J. Chem. Phys.* **2009**, *130*, 134114.

(94) Lindner, B.; Smith, J. C. *Comput. Phys. Commun.* **2012**, *183*, 1491-1501.