communications
biology

ARTICLE

https://doi.org/10.1038/s42003-021-01759-1 OPEN

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]✉

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. Here, we show that enhancing the sampling using Hamiltonian replica-exchange MD (HREMD) led to unbiased and accurate ensembles, reproducing small-angle scattering and NMR chemical shift experiments, for three IDPs of varying sequence properties using two recently optimized force fields, indicating the general applicability of HREMD for IDPs. We further demonstrate that, unlike HREMD, standard MD can reproduce experimental NMR chemical shifts, but not small-angle scattering data, suggesting chemical shifts are insufficient for testing the validity of IDP ensembles. 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). The results suggest that the major hurdle of generating an accurate unbiased ensemble for IDPs has now been largely overcome.

[1] Oak Ridge National Laboratory, Biosciences Division, UT/ORNL Center for Molecular Biophysics, Oak Ridge, TN, USA. [2] Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN, USA. ✉email: petridisl@ornl.gov

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-affinity interactions and recognition of partners by folding upon binding[1–6]. About 50–70% 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]. In addition, IDPs are an essential part of plant immune signaling components and also mediate plant–microbe interactions[7].

Understanding the function of a protein requires a determination of its 3D structure[8]. IDPs adopt highly dynamic structural ensembles, which are commonly characterized by nuclear magnetic resonance (NMR)[9], small-angle X-ray/neutron scattering (SAXS/SANS)[10,11], single-molecule fluorescence resonance energy transfer[12], hydrogen-exchange mass spectrometry[13], and circular dichroism[14,15]. However, the information content of the applied experimental techniques is insufficient to obtain the ensemble of 3D conformations an IDP adopts[16]. 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[2,17–28] have led to accurate simulation of short disordered peptides and model systems[18,19,29–32]. However, the simulations are not always consistent with the experiment, either because of inadequate sampling or shortcomings of the force fields[2,19,22,24,30,33,34].

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[35,36] or by a posteriori reweighting the ensemble of the MD population[37,38]. One challenge for these methods is degeneracy, that is, distinct 3D conformations may yield the same observable, which may lead to overfitting. Bayesian maximum entropy optimization approaches, which aim to perturb the MD ensemble as little as possible, have been employed to avoid overfitting[35,38,39]. 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[40,41]. HREMD improves sampling by scaling the intraprotein and protein-water potentials[17,20] of higher-order replicas, while keeping the potential of the lowest rank replica unscaled[42–45] so as to represent the physically meaningful interactions of the system. However, only two IDPs[40,41] 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[20] with TIP4P/2005s[20] and Amber ff99SB-*disp*[17] with modified TIP4P-D[17], 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 chemical shifts, 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 the 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 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 (Supplementary Note and Supplementary Fig. 1), employing two force fields: a03ws[20] and a99SB-disp[17]. Each replica of HREMD simulation is 500 ns long (Supplementary Tables 1–4). For comparison, we also conducted standard MD, that is, without enhancing the sampling, of the same cumulative length as the HREMD (Supplementary Tables 1–4). The cumulative lengths of standard MD simulations for Histatin 5, Sic 1, and SH4UD are 5, 8, and 10 μs, respectively. 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 Supplementary Fig. 2), 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 by the χ^2 value as defined in Eq. (5) and listed in Supplementary Table 5. χ^2 calculated from HREMD converges in ~100 ns for Histatin 5, but in ~300–400 ns for the larger Sic 1 and SH4UD (Supplementary Fig. 3).

The histograms of R_g show that standard MD simulations sample more compact structures than does HREMD with the same force fields. Moreover, the histograms of R_g from all the independent standard MD runs are different from each other (Supplementary Fig. 4), suggesting lack of convergence between the MD runs due to inadequate sampling. Therefore, for the IDPs studied here, poor agreement with experiment arises primarily from insufficient sampling rather than from shortcomings of the force fields.

NMR chemical shifts (CS) provide information on the local chemical environment of protein atoms and reflect structural factors such as backbone and side-chain conformations. The ensemble-averaged NMR secondary CS (ΔCS) calculated from all the simulations (force fields and sampling methods) are in a good agreement with the experimental values (Fig. 2a–c and Supplementary Figs. 5–8). The quality of agreement is determined from the root mean square (RMS) error defined in Eq. (6) for each backbone atom (Fig. 2d–f), which is of the order of predicted RMS errors of SHIFTX2[48] (viz, 1.12, 0.44, 0.52, 0.17, and 0.12 p.p.m. for N^H, $C^α$, $C^β$, H^N, and $H^α$ atoms, respectively[48]). However, a

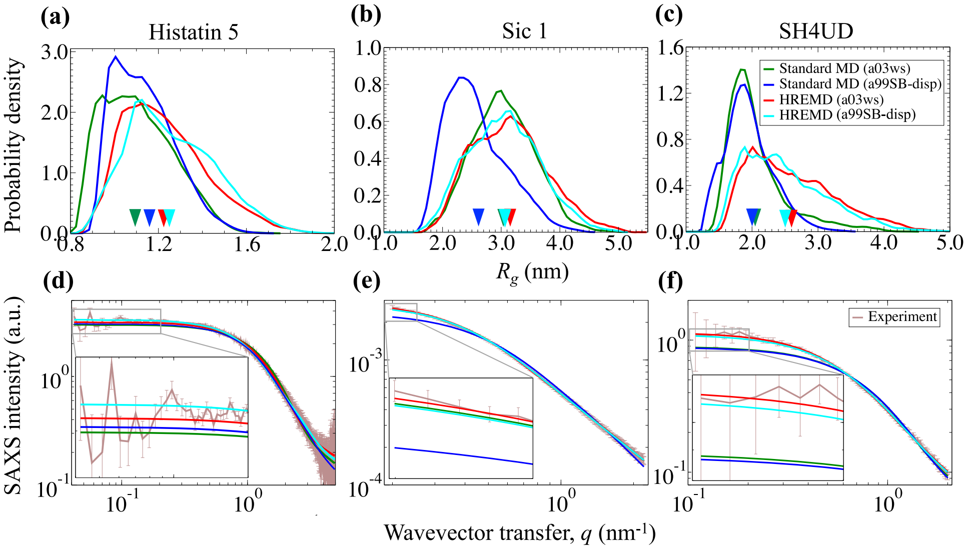

Fig. 1 Comparison of experiemntal and calculated global structural properties of IDPs. a–c The histograms of Rg of **a** Histatin 5, **b** Sic 1, and **c** SH4UD obtained from MD simulations. The inverted triangles indicate the average R_g of each simulation. **d–f** The SAXS profiles calculated from simulations (using SWAXS[46]) are compared to experiments for **d** Histatin 5[31], **e** Sic 1[47], and **f** SH4UD[41]. 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[20] or a99SB-disp[17], and sampling methods, standard MD or HREMD (Supplementary Tables 1 and 2). HREMD results are from the lowest rank replica of the simulations shown by the bold-italics font in Supplementary Table 2. SANS data of SH4UD are shown in Supplementary Fig. 2.

slightly better agreement between calculated and experimental ΔCS was obtained using a99SB-disp (Fig. 2d–f). Importantly, the agreement was not improved by enhancing the sampling by HREMD. Moreover, we found that the standard MD ensembles are consistent with NMR CS but not with small-angle scattering experiments. This indicates that CS alone may be an insufficient criterion to test the validity of IDP ensembles.

Both force fields and sampling methods predict nearly the same transient secondary structure elements. Transient helices, which are considered to be biologically relevant[51–53], were found proximal to known phosphorylation residues of Sic 1[47] and to known lipid-binding or phosphorylation residues in SH4UD[50,54]. In contrast, the propensity of each secondary structure element is found to depend on both the force fields and sampling methods (Supplementary Figs. 9 and 10). 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 (Supplementary Fig. 11).

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

$$C(s) = <\boldsymbol{n}_i \cdot \boldsymbol{n}_{i+s}> \tag{1}$$

where $s = |i - j|$ is the pairwise residue separation (sometimes called contour length), and $\boldsymbol{n}_i$ is the unit vector connecting C_α atoms of two consecutive residues (Fig. 3a). The steeper the decay of $C(s)$, the lower the stiffness of the chain. $C(s)$ is similar for the three IDPs for $s < 5$, exhibiting an exponential decay $C(s) = e^{-s/k}$, where k is the number of C_α atom pairs corresponding to the persistence length (l_p). 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. Here, we approximate $l_p = k \times 0.38$ nm, where 0.38 nm is the distance between two consecutive C_α atoms in proteins[55]. We obtain $k = 1.6$ and $l_p = 0.61 \pm 0.02$ nm for all IDPs, in close agreement to a value of $l_p = 0.40 \pm 0.07$ nm reported for unfolded (hCyp, CspTm, R15, and R17) and disordered (IN and ProTα, variants ProT53 and ProT54) proteins[55]. A power-law decay ($\sim s^{-3/2}$) is found for Sic 1 at $5 < s \leq 13$, whereas correlations decay more rapidly and vanish for $s > 5$ for Histatin 5 and SH4UD. Therefore, Sic 1 is the stiffest.

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

$$R_s = R_0 s^\nu \tag{2}$$

where R_s is the average intraprotein pairwise distance between the C_α atoms of residues i and j at 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 $\nu = 0.5$, while a self-avoiding random walk with $\nu = 0.588$ 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 ($s \leq 10$), all three IDPs show common behavior. Although the scaling law (Eq. (2)) is formally valid only for large chain lengths s, we fit the data for $s < 10$ with Eq. (2) to demonstrate that the IDPs have similar local chain statistics, evidenced by similar fits with $\nu \approx 0.70$ and prefactor of $R_0 \sim 0.4$ nm (R_0 is similar to the average distance between two consecutive C_α 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 self-avoiding random walk. 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[2,19,29,31,32], it has been necessary to bias or reweight the

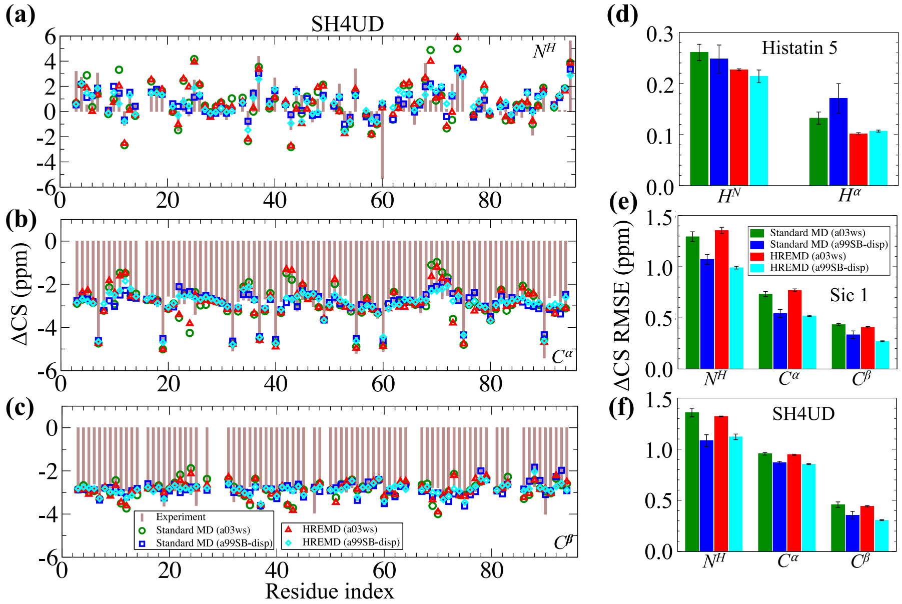

Fig. 2 Comparison of experimental and calculated local structural properties of IDPs. Comparison between the ensemble-averaged experimental (bars) and calculated (symbols) NMR secondary chemical shifts (ΔCS) of backbone atoms **a** N^H, **b** C^α, and **c** C^β for SH4UD. ΔCS RMSE of backbone atoms with respect to experimental values (bars), as defined in Eq. (6), for **d** Histatin 5[49], **e** Sic 1[47], and **f** SH4UD[50]. The error bars in ΔCS RMSE (**d**–**f**) are the standard error of the mean as defined in Eq. (4). The color code indicates the force field and sampling method used. The theoretical NMR chemical shifts are calculated using SHIFTX2[48]. The prediction values of SHIFTX2 have RMS errors of 1.12, 0.44, 0.52, 0.17, and 0.12 p.p.m. for backbone atoms N^H, C^α, C^β, H^N, and H^α, respectively[48].

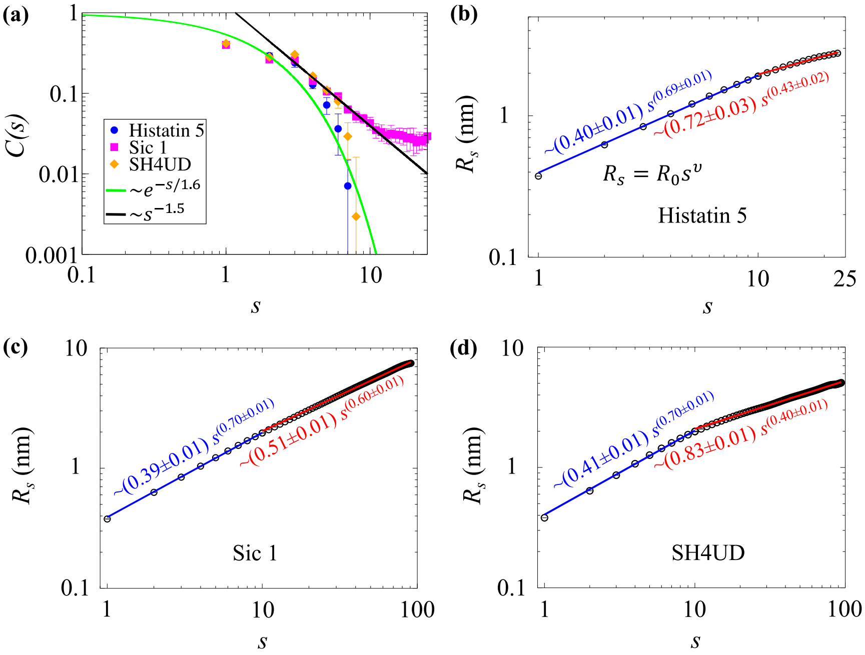

Fig. 3 Chain statistics of IDPs. **a** The orientational correlation function as a function of the pairwise residue sequence separation, s. For $s < 5$, $C(s)$ is fitted by $C(s) = e^{-s/k}$ for each IDP, where k is the number of C_α atom pair related to persistence length (l_p) by $l_p = k \times 0.38$ nm. For $s \geq 5$ 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_α 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, $s \leq 10$ (blue) and $s > 10$ (red). The error bars are smaller than the symbol size.

MD results to achieve consistency with experiments[35,36,38,39,61–63]. The reason MD has not always been accurate was unclear: it could have been 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 that the a03ws and a99SB-disp force fields are of adequate accuracy and that enhanced sampling techniques are necessary to reproduce the experimental data.

We found that the calculated NMR CS 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 (Supplementary Figs. 9 and 10), with a03ws and a99SB-disp having biases towards helices and β-sheets, respectively.

The data show that standard MD simulations can be in apparent agreement with NMR CS, 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 >~1 nm apart[61,65]) (Figs. 1 and 2 and Supplementary Fig. 2)[41]. Thus, agreement with NMR CS 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[17,66] of IDPs to ensure that 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[12,14,55,70,71]. The high fidelity HREMD trajectories reveal that, despite having markedly different sequences, the IDPs studied here have 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. 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 (Supplementary Fig. 1). 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 side chain, whereas the absence of a side chain for glycine increases the backbone flexibility[72,73], and therefore both are known to be disorder-promoting[72,73]. In comparison, SH4UD has both positively and negatively charged residues, 11 prolines and 12 glycines.

We now discuss the HREMD method[42,43,45] 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 scaling only the intraprotein and protein-solvent potential energy functions by a factor, λ and $\sqrt{\lambda}$, respectively (where $\lambda < 1$). An exchange of coordinates is allowed between neighboring replicas if the Monte Carlo metropolis criterion is satisfied[42,43]. 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, that is, HREMD 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 Supplementary Tables 2 and 3. A T_{max} of 400–450 K (lower limit) was needed to achieve a good agreement between the lowest rank (unscaled) replica of HREMD and the experimental SAXS results. Moreover, to estimate the upper limit of effective temperature, we performed HREMD of Histatin 5 using a99SB-disp, $T_{\mathrm{max}} = 800$ K and 24 replicas (Supplementary Table 3). This simulation generated the ensemble in the lowest rank replica similar to that of HREMD with $T_{\mathrm{max}} = 450$ K (Supplementary Fig. 12a). 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 (Supplementary Fig. 12a). However, choosing the higher value of T_{max} would increase the number of replicas and thus computational cost.

To ensure HREMD does not bias the ensemble, we also performed control simulations of a short intrinsically disordered peptide, Ala$_5$ (five residues)[17,20] and the 20-residue folded protein Trp cage[17,75] (Supplementary Fig. 13), both of which have been used as benchmarks for the optimization of molecular mechanics force fields[17,20]. Unlike what was observed for the longer IDPs (Histatin 5, Sic 1, and SH4UD), MD, and HREMD both yield similar ensembles for the controls (Supplementary Fig. 13). This suggests that HREMD does not introduce unphysical conformations and is equivalent to microsecond standard MD for short peptides and proteins.

A quantitative comparison of the sampling efficiency of standard MD and HREMD is provided by calculating the autocorrelation functions (C_{t}) of the number of contacts (n_{c}) and of R_{g} (Supplementary Note, Supplementary Fig. 14, and Supplementary Table 6). The decay of the autocorrelation is markedly more pronounced for HREMD than for standard MD. Taking the steepness of the decay of C_{t} in Supplementary Fig. 14 as a measure of sampling efficiency, it is clear that HREMD sampling is superior to that of standard MD (Supplementary Table 6).

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. The success of the HREMD approach for these three markedly different IDPs suggests that it will be of general applicability. 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 small-angle scattering experiments with standard MD of the same cumulative length as HREMD, although NMR CS were reproduced accurately with standard MD. Local chemical and structural properties of IDPs, which influence CS, therefore, seem force field-dependent, while the overall protein size and shape, which influences small-angle scattering intensities, also depend on the sampling. Both local and global features must be employed to validate IDP ensembles. Therefore, our results suggest adequately sampled simulations using recent IDP-specific force fields can reliably generate the 3D ensembles of IDPs (Supplementary Fig. 15), 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 (<~10 residues). More studies are required to establish whether this is a universal IDP behavior.

Methods

Experimental SAXS and NMR data. The experimental SAXS data of Histatin 5, Sic 1, and SH4UD were taken from Henriques et al.[31] Protein Ensemble Database (http://pedb.vib.be)[47] and our previous work[41], respectively. Similarly, NMR CS of backbone atoms (C^{α}, C^{β}, N^{H}, H^{α}, and H^{N}) of Histatin 5, Sic 1, and SH4UD were acquired from the literature[49], Protein Ensemble Database[47], and Biological Magnetic Resonance Data Bank database entry 15563[50] respectively.

MD simulations. The initial 3D structures of IDPs (Histatin 5, Sic 1, and SH4UD) were obtained from I-TASSER[76]. 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. A short disordered peptide, Ala_5[17,20], and a small folded protein, Trp-cage[17], were also simulated as controls (SI). The initial structure of Ala_5 was constructed using Visual MDs[77], whereas the starting structure of Trp cage was taken from PDB 1L2Y[75].

We performed standard MD simulations with two recently optimized force fields, Amber ff03ws[20,78,79] with TIP4P/2005s[20] (a03ws) and Amber ff99SB-*disp*[17,80] with the modified TIP4P-D[17,22] water model (a99SB-disp) using GROMACS[81–86]. All bonds involving hydrogen atoms were constrained using the LINCS algorithm[87]. The SETTLE algorithm was used for water[88]. 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[89] 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, 293, 300, 298, and 282 K for Histatin 5, Sic 1, SH4UD, Ala_5, and Trp cage, respectively, to match the temperatures measured at the experiments using velocity-rescaling thermostat[90] with a relaxation time of 0.1 ps. The pressure coupling was fixed at 1 bar using the Parrinello–Rahman algorithm[91] 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 (amount of substance, volume, and temperature) and NPT (amount of substance, pressure, and temperature) ensembles. The production runs were carried out in the NPT ensemble. The cumulative lengths of standard MD simulations with a number of independent runs enclosed in the brackets for Ala_5, Trp cage, Histatin 5, Sic 1, and SH4UD are 2 (1), 4 (4), 5 (5), 8 (4), and 10 µs (6), respectively (Supplementary Table 1).

Enhanced sampling MD simulations. We employed replica exchange with solute tempering 2[42,43], an HREMD simulation method to enhance the conformational sampling. Replica exchange with solute tempering 2 is implemented in GROMACS (v.2018.6)[81–86] patched with PLUMED (v.2.5.2)[92]. The interaction potentials of intraprotein and protein solvent were scaled by a factor λ and $\sqrt{\lambda}$, respectively, while water–water interactions were unaltered[42–44,93]. The scaling factor λ_i, and corresponding effective temperatures T_i of the *i*th replica are given by,

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

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 ($\lambda_0 = 1$ or T_0). Exchange of coordinate between neighboring replicas was attempted every 400 MD steps. Each replica of HREMD is 500 ns long. The cumulative lengths of HREMD simulations with the number of replicas enclosed in the brackets for Ala_5, Trp-cage, Histatin 5, Sic 1, and SH4UD are 2 (4), 4 (8), 5 (10), 8 (16), and 10 µs (20), respectively (Supplementary Tables 2 and 3). The details of HREMD and standard MD simulations are shown in Supplementary Tables 1–4. The secondary structure prediction was calculated with DSSP[94]. The orientational correlation function is determined using MD analysis[95].

Statistics and reproducibility. 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 ($i = 1$–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, and m_5) distribution, that is,

$$\text{Error bar} = \sqrt{\frac{1}{n(n-1)}\sum_{i}^{n=5}\left(m_i - \overline{m}\right)^2} \tag{4}$$

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

In regard to the reproducibility of the work, a multiple copies of standard MD and two copies of HREMD simulations were performed for each IDP.

Theoretical SAXS profiles. The theoretical SAXS and SANS intensities were calculated with SWAXS[46,96] and SASSENA[97], respectively, by taking into account of explicit hydration water, which contributes to the signal[46]. The agreement between experiment and simulation was determined by a χ^2 value:

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

where $<I_{\text{expt}}(q)>$ and $<I_{\text{sim}}(q)>$ 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. (5), 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[14] in the experiment.

Theoretical NMR CS. The theoretical NMR CS was calculated with SHIFTX2[48] by taking the average over all frames from the MD trajectory. Furthermore, we determined the NMR secondary CS for N^H, C^{α}, C^{β}, H^N, and H^{α} atoms as the difference between the experimental (or simulation-derived) CS and the corresponding random coil values specific to a particular atom and amino acid:

$$\Delta\text{CS}^{\text{expt}}(x,i) = \text{CS}^{\text{expt}}(x,i) - \text{CS}^{\text{RC}}(x,i)$$

$$\Delta\text{CS}^{\text{calc}}(x,i) = \text{CS}^{\text{calc}}(x,i) - \text{CS}^{\text{RC}}(x,i)$$

where atom $x \in N^H$, C^{α}, C^{β}, H^N, and H^{α} and i refers to a specific amino acid. $\text{CS}^{\text{expt}}(x,i)$, $\text{CS}^{\text{calc}}(x,i)$, and $\text{CS}^{\text{RC}}(x,i)$ are the CS values from experiment, MD (calculated) and random coil database for atom "x" and amino acid "i," respectively. Note that the calculated CS for each atom are corrected as $\text{CS}^{\text{calc}} = \text{CS}^{\text{calc0}} + O$, where CS^{calc0} is an actual ensemble-averaged value from SHIFTX2 and O is an offset constant determined from linear regression fit of the theoretical to the experimental NMR CS (Supplementary Figs. 5–7). Such an offset may arise from the systematic or referencing error in the NMR CS measurement or calculation[98] and is used here to improve the agreement between experiment and calculated values. CS^{RC} values are those reported in Tamiola et al.[99]. Finally, we quantified the agreement between experimental and calculated secondary CS of atom by evaluating the RMS error given by,

$$\text{RMSE}(x) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\Delta\text{CS}^{\text{expt}}(x,i) - \Delta\text{CS}^{\text{calc}}(x,i)\right)^2} \tag{6}$$

where n is the total number of residues in the IDP.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request. The source data for the Figs. 1, 2, and 3 are available as Supplementary Data 1, Supplementary Data 2, and Supplementary Data 3, respectively.

Code availability

The input files for running HREMD simulation using GROMACS patched with PLUMED are provided as Supplementary Data 4. It is also deposited in Zenodo[100] and GitHub (https://github.com/utsabstha/hremd-idp).

Received: 19 June 2020; Accepted: 7 January 2021;
Published online: 23 February 2021

References

1. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* **37**, 215–246 (2008).
2. Chong, S.-H., Chatterjee, P. & Ham, S. Computer simulations of intrinsically disordered proteins. *Annu. Rev. Phys. Chem.* **68**, 117–134 (2017).
3. Sun, X., Rikkerink, E. H., Jones, W. T. & Uversky, V. N. Multifarious roles of intrinsic disorder in proteins illustrate its broad impact on plant biology. *Plant Cell* **25**, 38–55 (2013).
4. Mitrea, D. M. et al. Nucleophosmin integrates within the nucleolus via multimodal interactions with proteins displaying R-rich linear motifs and rRNA. *eLife* **5**, e13571 (2016).
5. Martin, E. W. & Mittag, T. Relationship of sequence and phase separation in protein low-complexity regions. *Biochemistry* **57**, 2478–2487 (2018).
6. Shammas, S. L., Rogers, J. M., Hill, S. A. & Clarke, J. Slow, reversible, coupled folding and binding of the spectrin tetramerization domain. *Biophys. J.* **103**, 2203–2214 (2012).

7. Marín, M. & Ott, T. Intrinsic disorder in plant proteins and phytopathogenic bacterial effectors. *Chem. Rev.* **114**, 6912–6932 (2014).
8. van der Lee, R. et al. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
9. Dyson, H. J. & Wright, P. E. Perspective: the essential role of NMR in the discovery and characterization of intrinsically disordered proteins. *J. Biomol. NMR* **73**, 651–659 (2019).
10. Cordeiro, T. N. et al. Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Curr. Opin. Struct. Biol.* **42**, 15–23 (2017).
11. Mansouri, A. L. et al. Folding propensity of intrinsically disordered proteins by osmotic stress. *Mol. Biosyst.* **12**, 3695–3701 (2016).
12. Schuler, B., Soranno, A., Hofmann, H. & Nettels, D. Single-molecule FRET spectroscopy and the polymer physics of unfolded and intrinsically disordered proteins. *Annu. Rev. Biophys.* **45**, 207–231 (2016).
13. Balasubramaniam, D. & Komives, E. A. Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. *Biochim. Biophys. Acta* **1834**, 1202–1209 (2013).
14. Riback, J. A. et al. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science* **358**, 238 (2017).
15. Na, J. H., Lee, W. K. & Yu, Y. G. How do we study the dynamic structure of unstructured proteins: a case study on nopp140 as an example of a large, intrinsically disordered protein. *Int. J. Mol. Sci.* **19**, 381 (2018).
16. Baker, C. M. & Best, R. B. Insights into the binding of intrinsically disordered proteins from molecular dynamics simulation. *Wiley Interdiscip. Rev.* **4**, 182–198 (2014).
17. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl Acad. Sci. USA* **115**, E4758–E4766 (2018).
18. Huang, J. & MacKerell, A. D. Jr. Force field development and simulations of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **48**, 40–48 (2017).
19. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
20. 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.* **10**, 5113–5124 (2014).
21. Lindorff-Larsen, K., Trbovic, N., Maragakis, P., Piana, S. & Shaw, D. E. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc.* **134**, 3787–3791 (2012).
22. 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* **119**, 5113–5123 (2015).
23. Shabane, P. S., Izadi, S. & Onufriev, A. V. A general purpose water model can improve atomistic simulations of intrinsically disordered proteins. *J. Chem. Theory Comput.* **15**, 2620–2634 (2019).
24. 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* **59**, 1743–1758 (2019).
25. Yu, L., Li, D. W. & Bruschweiler, R. Balanced amino-acid-specific molecular dynamics force field for the realistic simulation of both folded and disordered proteins. *J. Chem. Theory Comput.* https://doi.org/10.1021/acs.jctc.9b01062 (2020).
26. Best, R. B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **42**, 147–154 (2017).
27. Best, R. B. Emerging consensus on the collapse of unfolded and intrinsically disordered proteins in water. *Curr. Opin. Struct. Biol.* **60**, 27–38 (2019).
28. 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.* **10**, 2227–2234 (2019).
29. Rauscher, S. et al. Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).
30. Henriques, J., Arleth, L., Lindorff-Larsen, K. & Skepo, M. On the calculation of SAXS profiles of folded and intrinsically disordered proteins from computer simulations. *J. Mol. Biol.* **430**, 2521–2539 (2018).
31. Henriques, J., Cragnell, C. & Skepo, M. Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. *J. Chem. Theory Comput.* **11**, 3420–3431 (2015).
32. Henriques, J. & Skepo, M. Molecular dynamics simulations of intrinsically disordered proteins: on the accuracy of the TIP4P-D water model and the representativeness of protein disorder models. *J. Chem. Theory Comput.* **12**, 3407–3415 (2016).
33. Lincoff, J., Sasmal, S. & Head-Gordon, T. The combined force field-sampling problem in simulations of disordered amyloid-beta peptides. *J. Chem. Phys.* **150**, 104108 (2019).
34. Bhowmick, A. et al. Finding our way in the dark proteome. *J. Am. Chem. Soc.* **138**, 9730–9742 (2016).
35. Hermann, M. R. & Hub, J. S. SAXS-restrained ensemble simulations of intrinsically disordered proteins with commitment to the principle of maximum entropy. *J. Chem. Theory Comput.* **15**, 5103–5115 (2019).
36. Roux, B. & Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **138**, 084107 (2013).
37. Fisher, C. K., Huang, A. & Stultz, C. M. Modeling intrinsically disordered proteins with Bayesian statistics. *J. Am. Chem. Soc.* **132**, 14919–14927 (2010).
38. Crehuet, R., Buigues, P. J., Salvatella, X. & Lindorff-Larsen, K. Bayesian-maximum-entropy reweighting of IDP Ensembles based on NMR chemical shifts. *Entropy* **21**, https://doi.org/10.3390/e21090898 (2019).
39. Hummer, G. & Kofinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **143**, 243150 (2015).
40. Liu, X. & Chen, J. Residual structures and transient long-range interactions of p53 transactivation domain: assessment of explicit solvent protein force fields. *J. Chem. Theory Comput* **15**, 4708–4720 (2019).
41. Shrestha, U. R. et al. Generation of the configurational ensemble of an intrinsically disordered protein from unbiased molecular dynamics simulation. *Proc. Natl Acad. Sci. USA* **116**, 20446–20452 (2019).
42. Wang, L., Friesner, R. A. & Berne, B. J. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J. Phys. Chem. B* **115**, 9431–9438 (2011).
43. Bussi, G. Hamiltonian replica exchange in GROMACS: a flexible implementation. *Mol. Phys.* **112**, 379–384 (2013).
44. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151 (1999).
45. Liu, P., Kim, B., Friesner, R. A. & Berne, B. J. Replica exchange with solute tempering: a method for sampling biological systems in explicit water. *Proc. Natl Acad. Sci. USA* **102**, 13749–13754 (2005).
46. Chen, P. C. & Hub, J. S. Validating solution ensembles from molecular dynamics simulation by wide-angle X-ray scattering data. *Biophys. J.* **107**, 435–447 (2014).
47. Mittag, T. et al. Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure* **18**, 494–506 (2010).
48. Han, B., Liu, Y., Ginzinger, S. W. & Wishart, D. S. SHIFTX2: significantly improved protein chemical shift prediction. *J. Biomol. NMR* **50**, 43–57 (2011).
49. Brewer, D., Hunter, H. & Lajoie, G. NMR studies of the antimicrobial salivary peptides histatin 3 and histatin 5 in aqueous and nonaqueous solutions. *Biochem. Cell Biol.* **76**, 247–256 (1998).
50. Pérez, Y., Gairí, M., Pons, M. & Bernadó, P. Structural characterization of the natively unfolded N-terminal domain of human c-Src kinase: insights into the role of phosphorylation of the unique domain. *J. Mol. Biol.* **391**, 136–148 (2009).
51. Kennedy, J. A., Daughdrill, G. W. & Schmidt, K. H. A transient alpha-helical molecular recognition element in the disordered N-terminus of the Sgs1 helicase is critical for chromosome stability and binding of Top3/Rmi1. *Nucleic Acids Res.* **41**, 10215–10227 (2013).
52. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signaling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
53. Hendus-Altenburger, R. et al. A phosphorylation-motif for tuneable helix stabilisation in intrinsically disordered proteins - lessons from the sodium proton exchanger 1 (NHE1). *Cell Signal.* **37**, 40–51 (2017).
54. Pérez, Y. et al. Lipid binding by the unique and SH3 domains of c-Src suggests a new regulatory mechanism. *Sci. Rep.* **3**, 1295 (2013).
55. Hofmann, H. et al. Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl Acad. Sci. USA* **109**, 16155–16160 (2012).
56. Valleau, J. P. Distribution of end-to-end length of an excluded-volume chain. *J. Chem. Phys.* **104**, 3071–3074 (1996).
57. Gomes, G.-N. W. et al. Conformational ensembles of an intrinsically disordered protein consistent with NMR, SAXS, and single-molecule FRET. *J. Am. Chem. Soc.* **142**, 15697–15710 (2020).
58. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384 (2014).
59. Wells, M. et al. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc. Natl Acad. Sci. USA* **105**, 5762–5767 (2008).
60. Uversky, V. N., Roman, A., Oldfield, C. J. & Dunker, A. K. Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs. *J. Proteome Res.* **5**, 1829–1842 (2006).
61. 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.* **129**, 5656–5664 (2007).
62. 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.* **138**, 094112 (2013).

63. Pitera, J. W. & Chodera, J. D. On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* **8**, 3445–3451 (2012).
64. Jensen, M. R., Zweckstetter, M., Huang, J. R. & Blackledge, M. Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem. Rev.* **114**, 6632–6660 (2014).
65. Neylon, C. Small angle neutron and X-ray scattering in structural biology: Recent examples from the literature. *Eur. Biophys. J.* **37**, 531–541 (2008).
66. Demerdash, O. et al. Using small-angle scattering data and parametric machine learning to optimize force field parameters for intrinsically disordered proteins. *Front. Mol. Biosci.* **6**, 64 (2019).
67. Shrestha, U. R. et al. Arabinose substitution effect on xylan rigidity and self-aggregation. *Cellulose* **26**, 2267–2278 (2019).
68. Baschnagel, J. et al. Semiflexible chains at surfaces: worm-like chains and beyond. *Polymers* **8**, https://doi.org/10.3390/polym8080286 (2016).
69. Ghosh, A. & Gov, N. S. Dynamics of active semiflexible polymers. *Biophys. J.* **107**, 1065–1073 (2014).
70. Fuertes, G. et al. Decoupling of size and shape fluctuations in heteropolymeric sequences reconciles discrepancies in SAXS vs. FRET measurements. *Proc. Natl Acad. Sci. USA* **114**, E6342–E6351 (2017).
71. Das, R. K. & Pappu, R. V. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl Acad. Sci. USA* **110**, 13392–13397 (2013).
72. Rauscher, S., Baud, S., Miao, M., Keeley, F. W. & Pomes, R. Proline and glycine control protein self-organization into elastomeric or amyloid fibrils. *Structure* **14**, 1667–1676 (2006).
73. Cheng, S., Cetinkaya, M. & Grater, F. How sequence determines elasticity of disordered proteins. *Biophys. J.* **99**, 3863–3869 (2010).
74. Nymeyer, H. How efficient is replica exchange molecular dynamics? An analytic approach. *J. Chem. Theory Comput.* **4**, 626–636 (2008).
75. Neidigh, J. W., Fesinmeyer, R. M. & Andersen, N. H. Designing a 20-residue protein. *Nat. Struct. Biol.* **9**, 425–430 (2002).
76. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
77. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
78. Duan, Y. et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012 (2003).
79. Best, R. B. & Mittal, J. Protein simulations with an optimized water model: cooperative helix formation and temperature-induced unfolded state collapse. *J. Phys. Chem. B* **114**, 14916–14923 (2010).
80. Hornak, V. et al. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **65**, 712–725 (2006).
81. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).
82. Lindahl, E., Hess, B. & van der Spoel, D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *Mol. Model. Annu.* **7**, 306–317 (2001).
83. Van Der Spoel, D. et al. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).
84. Suardíaz, R., Pérez, C., Crespo-Otero, R., García de la Vega, J. M. & Fabián, J. S. Influence of density functionals and basis sets on one-bond carbon−carbon NMR spin−spin coupling constants. *J. Chem. Theory Comput.* **4**, 448–456 (2008).
85. Pronk, S. et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
86. Abraham, M. J. et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
87. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
88. Miyamoto, S. & Kollman, P. A. Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comput. Chem.* **13**, 952–962 (1992).
89. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).
90. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
91. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
92. Bonomi, M. et al. PLUMED: a portable plugin for free-energy calculations with molecular dynamics. *Comput. Phys. Commun.* **180**, 1961–1972 (2009).
93. Peng, E., Todorova, N. & Yarovsky, I. Effects of forcefield and sampling method in all-atom simulations of inherently disordered proteins: application to conformational preferences of human amylin. *PLoS ONE* **12**, e0186219 (2017).
94. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
95. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32**, 2319–2327 (2011).
96. Park, S., Bardhan, J. P., Roux, B. & Makowski, L. Simulated x-ray scattering of protein solutions using explicit-solvent models. *J. Chem. Phys.* **130**, 134114 (2009).
97. Lindner, B. & Smith, J. C. Sassena - X-ray and neutron scattering calculated from molecular dynamics trajectories using massively parallel computers. *Comput. Phys. Commun.* **183**, 1491–1501 (2012).
98. Marsh, J. A., Singh, V. K., Jia, Z. & Forman-Kay, J. D. Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: implications for fibrillation. *Protein Sci.* **15**, 2795–2804 (2006).
99. Tamiola, K., Acar, B. & Mulder, F. A. A. Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J. Am. Chem. Soc.* **132**, 18000–18003 (2010).
100. Shrestha, U. R., Smith, J. C. & Petridis, L. Input files and scripts for Hamiltonian replica-exchange molecular dynamics simulations of intrinsically disordered proteins using a software GROMACS patched with PLUMED [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.4319228 (2020).

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 under an ALCC award 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. 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).

Author contributions

U.R.S. and L.P. conceived and designed research; U.R.S. performed research; U.R.S. and L.P. analyzed the data; J.C.S. discussed results and edited the manuscript; U.R.S. and L.P. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-021-01759-1.

Correspondence and requests for materials should be addressed to L.P.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.