Electronic Supplementary Information: Sequence Effects on Size, Shape, and Structural Heterogeneity in Intrinsically Disordered Proteins

Upayan Baul,[†,¶] Debayan Chakraborty,[†] Mauro L. Mugnai,[†] John E. Straub,[‡] and D. Thirumalai[*,†]

†*Department of Chemistry, The University of Texas at Austin, Austin, Texas 78712*

‡*Department of Chemistry, Boston University, Boston, Massachusetts 02215*

¶*Current address: Institute of Physics, Albert-Ludwigs-University of Freiburg, Hermann-Herder-Strasse 3, 79104 Freiburg, Germany*

E-mail: dave.thirumalai@gmail.com

Development of the SOP-IDP model

We set ourself the task of creating a minimal model, which could be used for IDPs with arbitrary length and sequence, in order to accurately describe not only the average properties of IDPs but also the details of their structural ensemble. Although not investigated here, the model ought to be simple enough to include the effects of denaturants, as is often done in experiments. To this end, we created the SOP-IDP model, which is built on the successful Self-Organized Polymer (SOP) model used to study temperature and denaturant dependent folding thermodynamics and kinetics of a large number of globular proteins.[1,2] Unlike in the case of globular protein folding, where neglecting non-native interactions may be justified,[3,4] in describing IDPs all amino residues have to be treated on equal footing. In the SOP-IDP model each residue is represented by a C_α atom, and a side-chain (SC) bead that is covalently bonded to the C_α atom. Exceptions to this representation are glycine and alanine, which are represented in the SOP-IDP model by single beads owing to their small sizes. In the implementation of the pair potentials, the glycine and alanine beads are thus treated as both side-chain beads, and backbone beads, depending on the type of the partner. The interactions between pairs of glycine or alanine beads are treated through the SC-SC interaction potential in the energy function, to account for sequence specificity. The charges and the van der Waals radii, which are needed for integrating the low friction equations of motion using the SOP-IDP force field, for all the interaction sites are given in Table S1.

Table S1: Parameters for the coarse-grained beads in the SOP-IDP model.

bead type	vdW radius ($\mathring{A}$)	charge (e)
C_α	1.90	0.0
Gly	2.25	0.0
Ala	2.52	0.0
Arg	3.28	1.0
Lys	3.18	1.0
His	3.04	0.0 / 1.0
Asp	2.79	-1.0
Glu	2.96	-1.0
Ser	2.59	0.0
Thr	2.81	0.0
Asn	2.84	0.0
Gln	3.01	0.0
Cys	2.74	0.0
Pro	2.78	0.0
Ile	3.09	0.0
Leu	3.09	0.0
Met	3.09	0.0
Phe	3.18	0.0
Trp	3.39	0.0
Tyr	3.23	0.0
Val	2.93	0.0

Learning procedure for parametrizing the SOP-IDP model

The pre-factors ϵ_{BB}, ϵ_{BS}, and ϵ_{SS}, which set the energy scales corresponding to the non-local interactions (see eq 1 in the main text), are the only free parameters in the SOP-IDP energy function. We used the experimental estimates of R_g as well as the low q regions of the SAXS profiles for three relatively short IDP sequences, Histatin-5, ACTR, and hNHE1, to obtain the initial estimates for the three free parameters, ϵ_{BB}, ϵ_{BS}, ϵ_{SS} in the SOP-IDP model.[5,6] Histatin-5 is a small (24 residue) IDP that has often been used as a reference for testing the validity of computational models.[5,7] The other two IDPs, ACTR ($N_T = 71$) and hNHE1 ($N_T = 131$) are also well studied.[6,8] In addition, the SAXS profiles for Histatin-5, ACTR and hNHE1 at 150 mM monovalent salt concentration have low noise-to-signal ratio, which makes objective comparisons feasible with the simulated profiles at equivalent ionic strengths.

However, Histatin-5, ACTR, and hNHE1 are short compared to other frequently studied IDP sequences. Thus, we expanded our training set, and we refined the SOP-IDP energy function by using experimental R_g values for the K32, K23, and hTau40 sequences. The optimal set of parameters for the SOP-IDP model from our learning procedure is:

$$\epsilon_{BB} = 0.12\,\text{kcal/mol} \quad \epsilon_{BS} = 0.24\,\text{kcal/mol}, \quad \epsilon_{SS} = 0.18\,\text{kcal/mol}. \tag{1}$$

which in units of k_BT (T=298 K) are $\epsilon_{BB} = 0.2$, $\epsilon_{BS} = 0.4$, $\epsilon_{SS} = 0.3$. The values quoted in eq 1 are different from the ones used for globular proteins.[1] The maximum decrease is in ϵ_{BB}, which is 4.6 times smaller than used previously, whereas ϵ_{BS} and ϵ_{BS} are only ≈ 1.7 times smaller (see Table S1 in Liu *et al.*[1]). Because the largest change is in ϵ_{BB}, we recalculated R_g for Histatin-5 using the value for ϵ_{BB} used previously.[1] The resulting value is only about 16% smaller than what is reported in Table S3.

Our rationale for seeking a different set of parameters are summarized below. (1) The values for globular proteins describe the situation with no denaturants (concentration of denaturants, $[C] = 0$) so that the folded states are predominantly populated. As $[C]$ increases, the stabilities of the folded states decrease, which we accounted for phenomenologically using transfer free energies, thus creating the SOP-MTM model.[9,10] Within this framework, the relevant interaction energy scales would decrease linearly. For example, $\epsilon_{BS} \approx \epsilon_{BS}([C] = 0) - m_{BS}[C]$ where m_{BS} is the analogue of the m value accounting for the loss of global stability at non-zero $[C]$. (2) Because the ensemble of conformations of IDPs behave like the unfolded states of globular proteins, created at high denaturant concentrations, we reasoned that it is not appropriate to use the values that stabilize the folded states.

The current SOP-IDP model describes IDPs and denatured states of globular proteins. It cannot describe with near quantitative accuracy the states of globular proteins at all values of denaturant concentration. We hasten to add that using the SOP-IDP parameters in eq 1, we find that the mean R_g of the hairpin from the GB1 protein is $\approx 1.04\,\text{nm}$, which compares favorably with the value ($R_g \approx 1.22\,\text{nm}$) calculated using the PDB structure. However, we should emphasize that there is no guarantee that the SOP-IDP model could describe the fate of globular proteins of arbitrary size and topology, at various external conditions, as accurately as we have done here for IDPs. Indeed, currently, no such force field at any level of description exists that can achieve this goal, and it is unlikely that a universal force field could be constructed, which would be accurate (errors in directly measurable quantities, when compared to experiments, that are small for a number of systems over a range of external conditions) for both globular proteins and IDPs. Construction of such a force field would be equivalent to solving the protein folding problem (prediction of structure, thermodynamics, and kinetics) from sequence alone.

Quantitative comparison between the simulated and experimental SAXS profiles

The level of agreement between the simulated and experimental SAXS profiles (in the Kratky representation) for the 24 IDP sequences (whose scattering profiles are depicted in Fig. 1 and Fig. 2 in the main text) could be quantified by calculating the extent of deviation between the simulated and experimental SAXS profiles. From the simulation trajectories, the $q^2 I_q/I_0$ values were computed for each IDP sequence at discrete intervals of q ($\Delta q = 0.01\mathring{A}^{-1}$). For a quantitative comparison, the number of data points available from the experimental SAXS profiles were reduced to produce datasets equivalent to those obtained from simulations.

Following the discretization, we calculated,

$$\tilde{\delta}^2 = \sum_{q_i \leq (qR_g)_{max}} (X_{i,exp} - X_{i,sim})^2 \tag{2}$$

where $X_i = (q_i)^2 I_{q_i}/I_0$ and $[0,(qR_g)_{max}]$ is the range over which $\tilde{\delta}^2$ is estimated. In order to compare the relative errors for all the IDPs on equal footing, we report the error estimates in terms of δ^2, which is defined as:

$$\delta^2 = \frac{\tilde{\delta}^2}{M}, \tag{3}$$

M being the number of data points available for comparison between the experimental and simulated data for a given IDP and for a given choice of $(qR_g)_{max}$.

The δ^2 estimates corresponding to the simulated SAXS profiles (shown in Figs 1 and 2 in the main text) are tabulated in Table S2 for two different choices of $(qR_g)_{max}$. Table S2 shows that for values of $(qR_g)_{max}$ up to 3 (well beyond the normal Guinier regime) the relative errors are small while they increase at large $(qR_g)_{max}$. The results in Table S2 show that the predictions of the SOP-IDP model are fairly accurate.

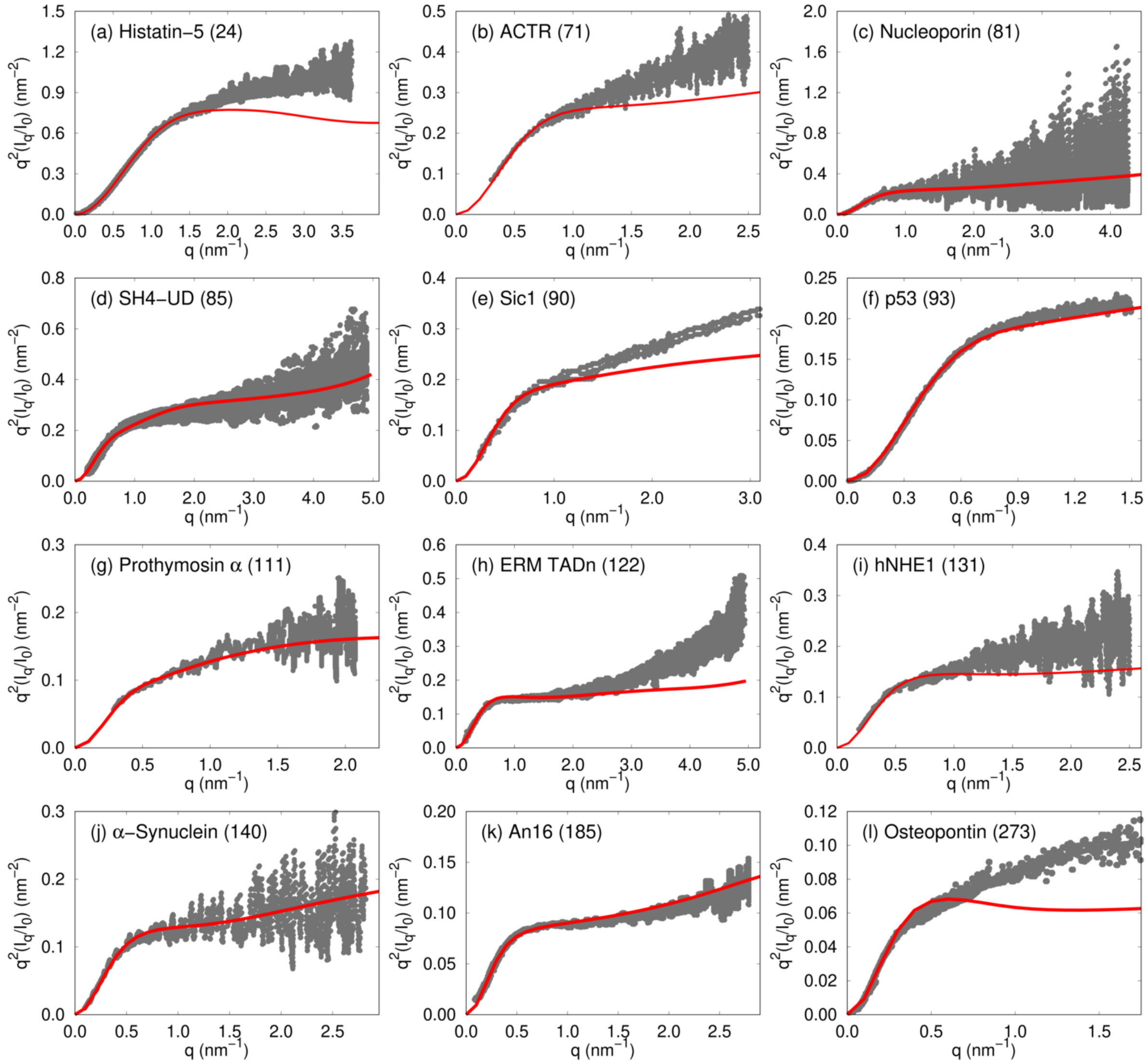

Figure S1: The Kratky plots for twelve IDP sequences. The values of N_T are in the parentheses. The gray points denote the experimental data, and the red curves denote the simulated profiles. In almost all cases, the agreement between the simulations and experiment is excellent in the small q region.

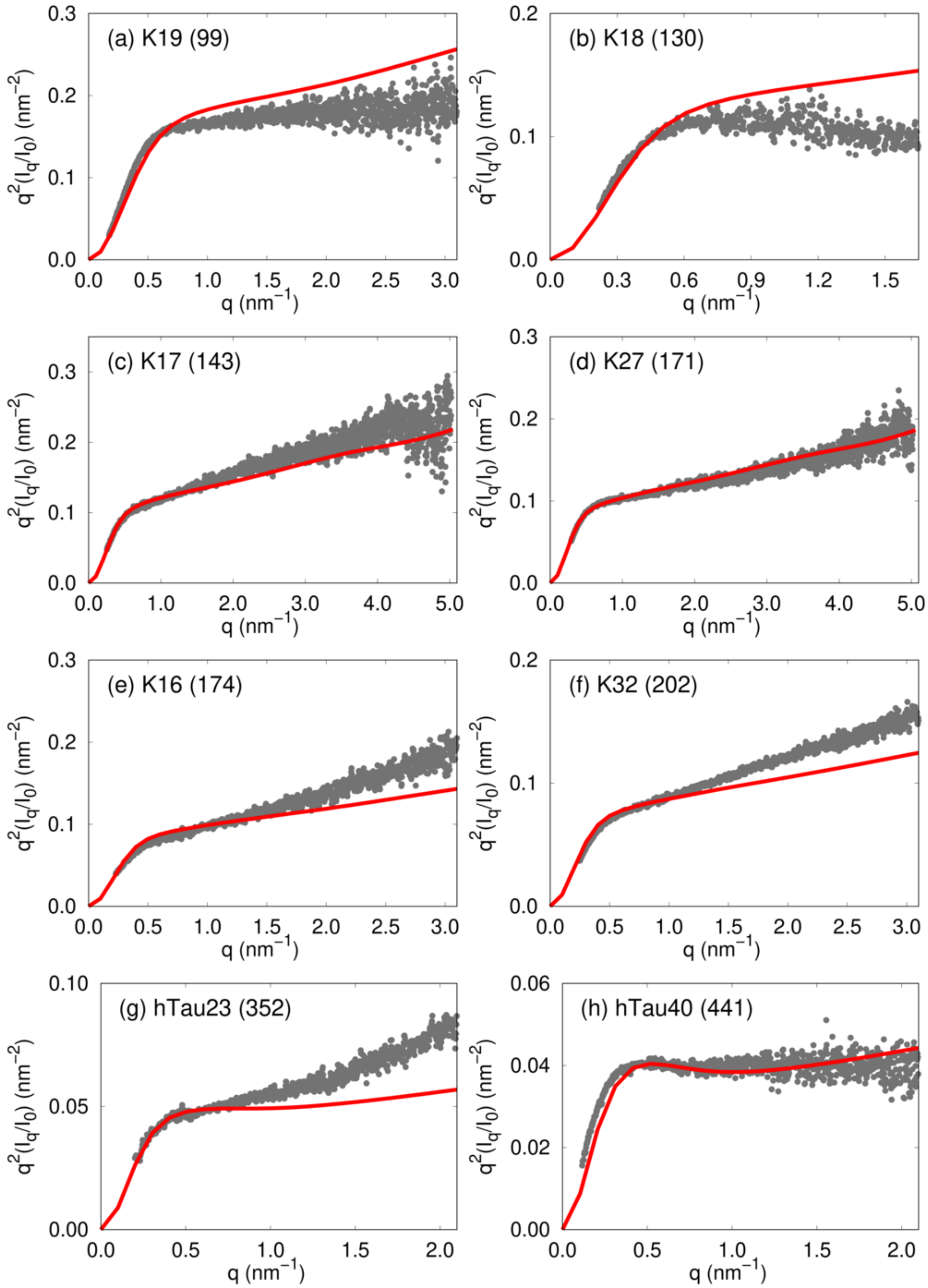

Figure S2: The Kratky plots for the Tau IDP sequences. The gray points denote the experimental data, and the red curves denote the simulated profiles. As in Fig. S1, the values of N_T are given in parentheses.

Table S2: Calculated δ^2 between experimental and simulated SAXS profiles. The numbers in parentheses (M) represent the number of data points compared for the respective estimates.

IDP	$(\delta^2\|_{(qR_g)_{max}=3.0} \times 10^{-4})$ (M)	$(\delta^2\|_{(qR_g)_{max}=5.0} \times 10^{-4})$ (M)
Histatin-5	6.81 (22)	79.42 (36)
ACTR	2.08 (9)	10.22 (17)
Nucleoporin	1.81 (11)	3.12 (19)
SH4-UD	1.87 (9)	0.91 (17)
Sic1	1.73 (8)	2.54 (15)
p53	0.18 (10)	0.41 (14)
Prothymosin α	0.11 (5)	0.70 (10)
ERM TADn	2.05 (8)	3.51 (15)
hNHE1	1.26 (7)	2.18 (13)
α-Synuclein	0.65 (7)	0.81 (13)
An16	0.19 (5)	0.34 (9)
Osteopontin	0.53 (10)	3.72 (14)
K19	1.02 (9)	3.37 (15)
K18	1.13 (6)	5.97 (12)
K17	0.01 (5)	0.08 (11)
K27	0.18 (5)	0.16 (10)
K16	0.05 (5)	0.10 (10)
K32	0.03 (4)	0.07 (9)
hTau23	0.02 (4)	0.03 (7)
hTau40	0.08 (3)	0.05 (6)

Comparison between simulated and experimental R_g

In contrast to SAXS experiments, where R_g values are usually determined from a Guinier analysis of the scattering profiles in the low q regime, we can obtain R_g directly from simulations, using the standard polymer physics formula:

$$R_g = \sqrt{\frac{1}{N}\langle\sum_{i=1}^{N}(r_i - r_{CM})^2\rangle} \tag{4}$$

In eq 4, N denotes the number of beads, r_i are the coordinates of bead i, r_{CM} is the centre-of-mass coordinate, and $\langle ... \rangle$ denotes the ensemble average.

The correlation plot shown in Fig. S3 quantifies the high level of agreement between the experimental and simulated R_g values. The corresponding % errors in our prediction are quoted in Table S3. Using an expression similar to eq 3, we find that the relative error (δ^2) between the experimental and simulated R_g values is 0.13.

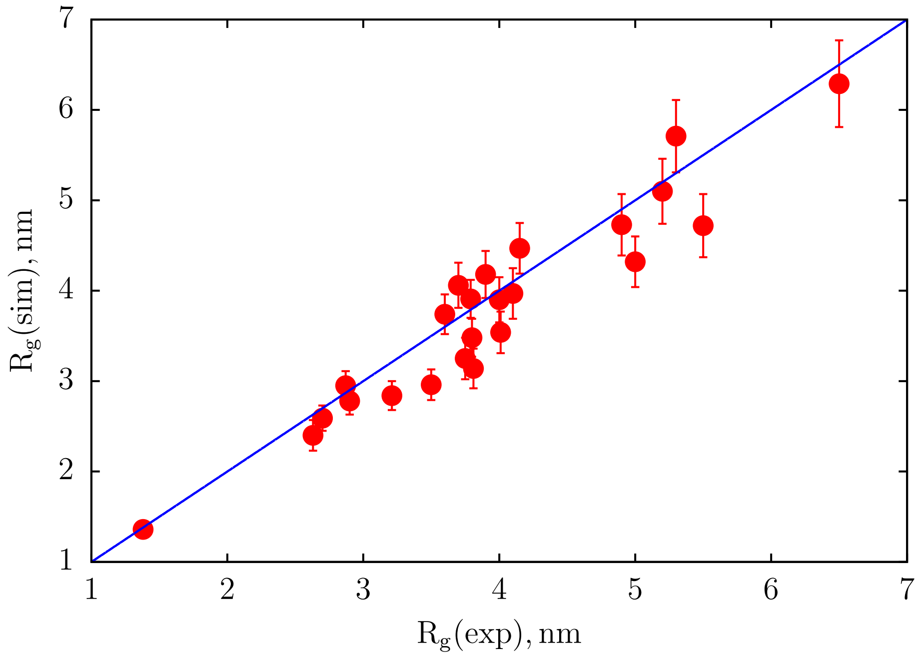

Figure S3: Comparison between the experimental R_g values and those obtained from simulations for the 24 IDP sequences. The blue solid line, which provides a guide to the eye, shows excellent agreement, especially when considering errors in experiments.

Table S3: Conformational properties of IDPs. The values of R_g, R_{ee} and R_h are in nm. The numbers in parenthesis represent standard errors. Column 5 denotes the % error (=100 $\frac{|R_g(exp)|-|R_g(sim)|}{R_g(exp)}$) in our prediction of R_g values for the different IDPs. The reasons for large errors in the mean Δ and S are explained in the SI text. [a] denotes the experimental value of R_h for the Sic1 sequence at 150 mM reported in Liu *et al.*[11] [b] denotes the experimental R_h value for Prothymosin-α reported by Uversky *et al.*[12] [c] denotes the experimental R_h value for α-Synuclein estimated by Paleologou *et al.*[13]

IDP	N_T	R_g (exp)	R_g (sim)	% error	R_{ee} (sim)	R_h (sim)	Δ	S
Histatin-5	24	1.38	1.36	1.5	3.09	1.32	0.40 (0.16)	0.50 (0.34)
ACTR	71	2.63	2.40	8.7	5.62	2.07	0.41 (0.19)	0.49 (0.39)
Nucleoporin	81	2.7	2.59	4.0	6.18	2.19	0.40 (0.19)	0.49 (0.41)
SH4-UD	85	2.9	2.78	4.1	6.55	2.30	0.43 (0.19)	0.53 (0.43)
Sic1	90	3.21	2.84	11.5	6.77	2.36 (2.2)[a]	0.42 (0.19)	0.52 (0.40)
p53	93	2.87	2.95	2.8	7.08	2.43	0.42 (0.19)	0.54 (0.40)
Prothymosin α	111	3.79	3.91	3.2	9.37	2.92 (3.1)[b]	0.51 (0.19)	0.72 (0.45)
ERM TADn	122	3.81	3.14	17.5	7.37	2.63	0.38 (0.19)	0.45 (0.39)
hNHE1	131	3.75	3.25	13.3	7.45	2.69	0.39 (0.19)	0.48 (0.39)
α-Synuclein	140	4.00	3.54	11.7	8.16	2.89 (2.82)[c]	0.40 (0.19)	0.47 (0.40)
An16	185	5.0	4.32	13.5	10.24	3.38	0.43 (0.19)	0.54 (0.41)
Osteopontin	273	5.5	4.72	14.2	10.08	3.82	0.40 (0.19)	0.52 (0.40)
K19	99	3.5	2.96	15.4	7.00	2.44	0.43 (0.19)	0.54 (0.41)
K18	130	3.8	3.48	8.4	8.24	2.80	0.43 (0.19)	0.54 (0.42)
K17	143	3.6	3.74	3.8	8.89	2.98	0.43 (0.19)	0.54 (0.42)
K10	167	4.0	3.90	2.5	9.08	3.13	0.41 (0.19)	0.50 (0.41)
K27	171	3.7	4.06	9.7	9.64	3.20	0.43 (0.19)	0.55 (0.42)
K16	174	3.9	4.18	7.2	9.89	3.28	0.42 (0.19)	0.51 (0.40)
K25	185	4.1	3.97	3.2	9.03	3.22	0.36 (0.18)	0.41 (0.37)
K32	202	4.15	4.47	7.7	10.57	3.49	0.42 (0.19)	0.52 (0.42)
K23	254	4.9	4.73	3.5	10.75	3.75	0.38 (0.19)	0.44 (0.39)
K44	283	5.2	5.10	1.9	11.63	4.01	0.40 (0.19)	0.49 (0.40)
hTau23	352	5.3	5.71	7.7	12.79	4.44	0.42 (0.19)	0.53 (0.40)
hTau40	441	6.5	6.29	3.2	14.11	4.88	0.42 (0.19)	0.54 (0.41)

Comparisons with results from all-atom force fields

In this section, we compare SOP-IDP simulation results for the SAXS profiles to available results obtained using simulations based on atomically detailed force fields. In the last few years, atomic detailed simulations, with vastly different force fields, have been used to calculate the SAXS profiles for the 71-residue ACTR[8] and the 24-residue RS peptide (FASTA sequence GAMGPSYGRSRSRSRSRSRSRSRS).[14] For ACTR,[8] the simulated $I(q)$ and the radius of gyration (R_g), which were calculated by adjusting the Lennard-Jones interaction strength between the oxygen atom of water and the heavy atoms on the protein within the AMBER force field, gave improved results relative the the Amber ff03 force field. As noted by the authors[8] the overall dimension of ACTR is still more compact relative to the experimentally measured R_g. However, simulations using the same force field as before[8] performed in a larger simulation box have lead to $I(q)$ (Best, personal communication) for ACTR that is in as excellent agreement with predictions based on SOP-IDP simulations for ACTR. (see Fig. 1b in the main text).

The experimental measurements of the SAXS profile for the RS peptide[15] have been used to compare and validate the state-of-the-art all-atom force fields.[14,15] Using an entirely different force field (TIP4P-D water model, adjustment of residue specific parameters, and addition of corrections to H bond interactions) Wu, Jiang, and Wu (WJW) arrived at the RSFF2+/TIP4P-D potential, here referred to as the WJW model. The WJW force field was used to calculate the SAXS profiles for the 24 residue Histatin-5 (see Fig. 1a in the main text) and the RS peptide. Comparison of these results with experiments showed excellent agreement with experimental $I(q)$ (see Fig. S7 in Wu *et al.*[14]). Note that the force fields used in these studies[8,14–16] are very different. For example, both in[14,16] the H-bond between the backbone carbonyl oxygen and amide hydrogen is unusually short (0.15 nm versus the usual ~ 0.2 nm in AMBER force fields) and the corresponding Lennard-Jones interaction is considerably stronger (~ 0.3 kcal/mole versus 0.057 kcal/mole in AMBER). There are many other parameters that also vary greatly between the different force fields.

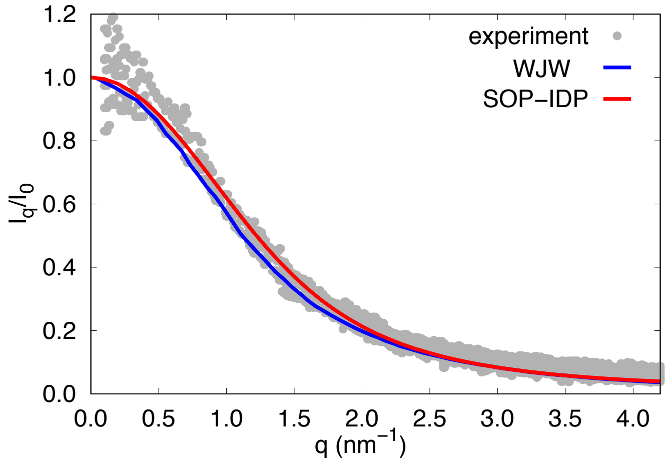

Figure S4: Top: Comparison of simulated SAXS profiles for the RS peptide using the SOP-IDP model with experiment,[15] and MD simulations by Wu *et. al.* (RSFF2+).[14]

In the following, we briefly compare results from simulations using the SOP-IDP model for the RS peptide, with results from all-atom simulation models, and experiment. The results of the SAXS experiments for the RS peptide were reported by Grubmüller *et. al.* at 298 K, neutral pH, 100 mM NaCl, and 50 mM Na-phosphate buffer. The experimental value of R_g is 1.262±0.007 nm.[15] They also benchmarked the performances of multiple state-of-the-art peptide force fields including AMBER, OPLS, and CHARMM variants (for details please refer to Table 1 in Grubmüller *et al.*[15]) against the experimental observations (Recent comparison of the performances of different force fields may be found in[14,16]). The large-scale simulations, using replica-exchange molecular dynamics at temperatures between 298 K and 450 K, were performed at 150 mM NaCl concentration.[15] The best agreement with the experimental R_g was obtained with the CHARMM 22* force field (1.265±0.007 nm) in conjunction with CHARMM-modified TIP3P water. The same peptide force field when used in conjunction with the dispersion-corrected TIP4P-D water model,[14,17] however resulted in a somewhat larger R_g of ∼1.4 nm. Several other combinations of peptide and water force fields yielded a wide range of R_g values ∼(1.0–1.5) nm.[15] The RS peptide was also used to validate the WJW force field,[14] which was optimized based on the AMBER-ff99SB but with a different water model and adjustments to many other parameters. The R_g value was obtained to be 1.32±0.005 nm, also using replica-exchange molecular dynamics simulations, and at a salt concentration adequate for neutralization of the peptide charges.

Using the SOP-IDP model at 150 mM salt concentration, we obtained the R_g value of 1.293±0.07 nm for the RS peptide, which is in excellent agreement (a mere 2.5 % difference) with the experimental measurement. In Fig. S4 the SAXS profiles obtained from our simulations of the RS peptide using the SOP-IDP model is compared with the experimental SAXS profile,[15] and simulated $I(q)$ using the WJW force field.[14] The latter data are extracted from the corresponding references using the online tool *WebPlotDigitizer*.[18] As can be clearly seen, simulated profiles from both the models are within the dispersion of experimental data. This comparison also illustrates that one could construct several force fields (both atomically detailed as well as coarse-grained), which could yield comparable results for small IDPs. The challenge is to *predict* results for IDPs of arbitrary length and sequences. Comparison with experiments will ultimately be the sole test of accuracy.

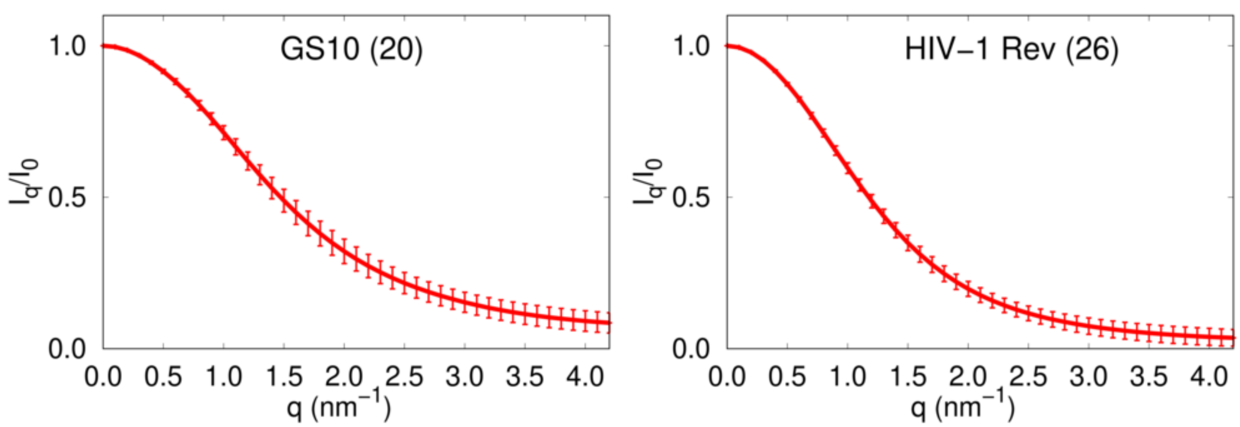

Figure S5: Predicted SAXS profiles for the GS10 peptide and the argininine rich motif of HIV-1 Rev (see text for sequences and simulation conditions) using the SOP-IDP model. The numbers in parentheses denote sequence lengths.

Predictions using the SOP-IDP model

We also extended the SOP-IDP simulations to two additional disordered sequences, the charge-neutral GS10 (FASTA sequence GSGSGSGSGSGSGSGSGSGS) and the arginine rich motif of HIV-1 Rev (FASTA sequence GAMATRQARRNRRRRWRERQRAAAAR). To the best of our knowledge, experimental SAXS profiles are not available for these peptides. For these sequences, WJW reported reported the R_g values $\sim 0.936 \pm 0.05$ nm for GS10 and $\sim 1.49 \pm 0.04$ nm for HIV-1 Rev[14] (data extracted digitally from the pink colored bar graph in Figure S6 in Wu *et al.*[14]). The corresponding values from simulations using the SOP-IDP model at 298 K are 1.027 ± 0.08 nm and 1.465 ± 0.096 nm for GS10 and HIV-1, respectively. In these two cases the R_g values predicted using simulations with the state-of-the-art WJW force field and the SOP-IDP simulations are in excellent agreement. The simulations for the HIV-1 Rev peptide using the SOP-IDP model were performed at an effective salt concentration of 10 mM (simulations by Wu *et al.* were reported at a salt concentration adequate for neutralization of the peptide charges). In Fig. S5 we plot the simulated SAXS profiles. For both these peptides the calculated $I(q)$ profiles serve as testable predictions.

A note concerning achieving good agreement between simulations and experiments is in order. For the three small IDPs (20 residue GS10, 24 residue RS peptide, and the 26 residue HIV-1 Rev) the R_g values can be readily calculated using the fit $R_g = 0.2N^{0.588}$ given in the caption to Fig. 3a in the main text. This Flory formula gives R_g values of 1.164 nm, 1.296 nm, and 1.358 nm, for GS10, RS peptide, and HIV-1 Rev respectively. These theoretical results do not differ significantly from the simulation results quoted above, which shows that in order to assess the accuracy of the force fields, at any level of coarse-graining, for IDPs or globular proteins one has to compare predictions with directly measurable quantities, such as $I(q)$ as a function of q and heat capacity.

Distributions of shape parameters

In Figures S6–S9, we show the distributions of shape parameters Δ and S (see eq 5 in the main text) for all the IDP sequences. Remarkably, the results in Figures S6–S9 show that $P(\Delta)$ and $P(S)$ are broad, with large dispersions. This is due to the plasticity of conformations sampled by the IDPs, which is further corroborated using hierarchical clustering of the IDP conformations (as discussed in the main text). The large dispersions result in substantial standard errors (Table S3), which merely show that the average Δ and S do not elucidate the granularity of the IDP conformations.

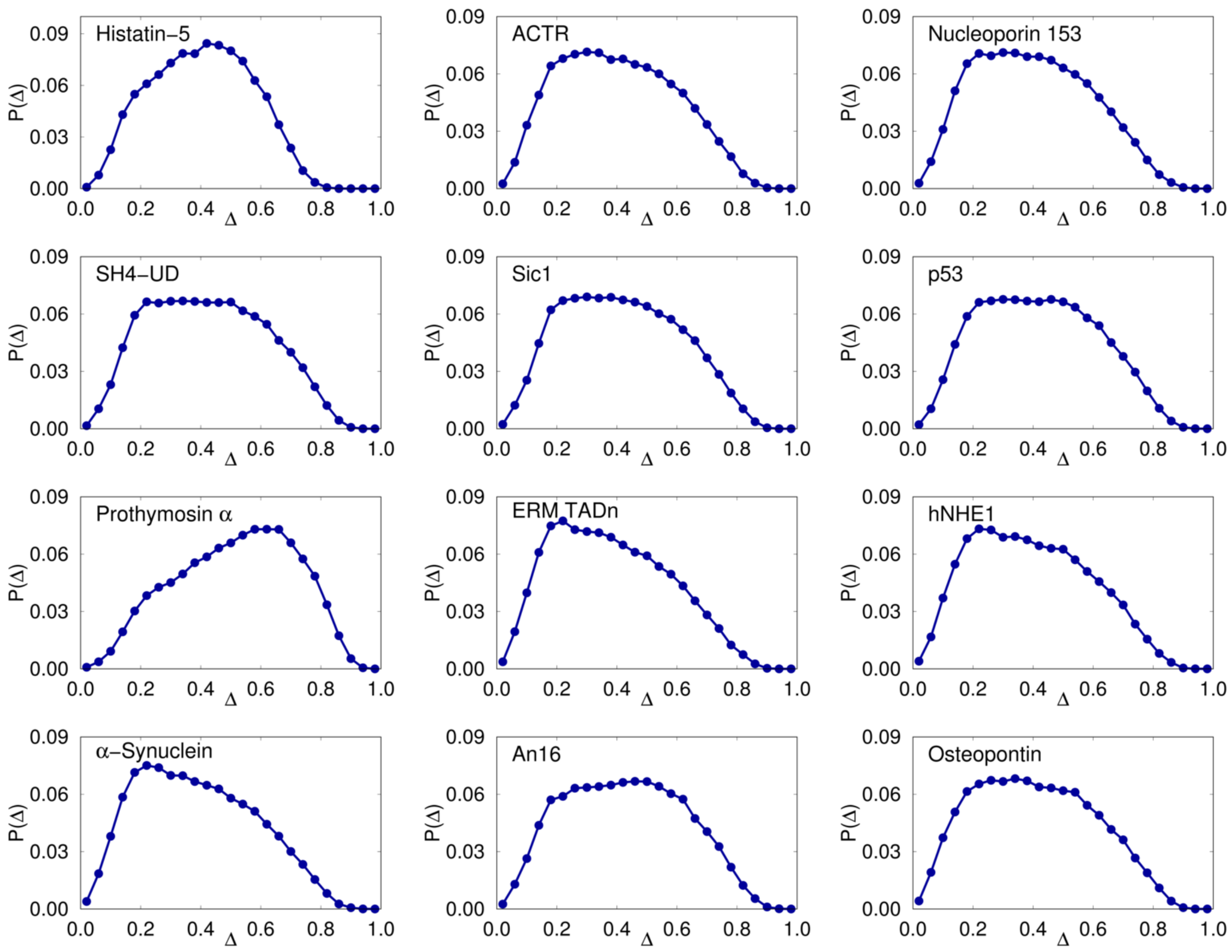

Figure S6: Distributions of the shape parameter Δ (eq 5 in the main text), for all non-Tau IDP sequences listed in Table S3.

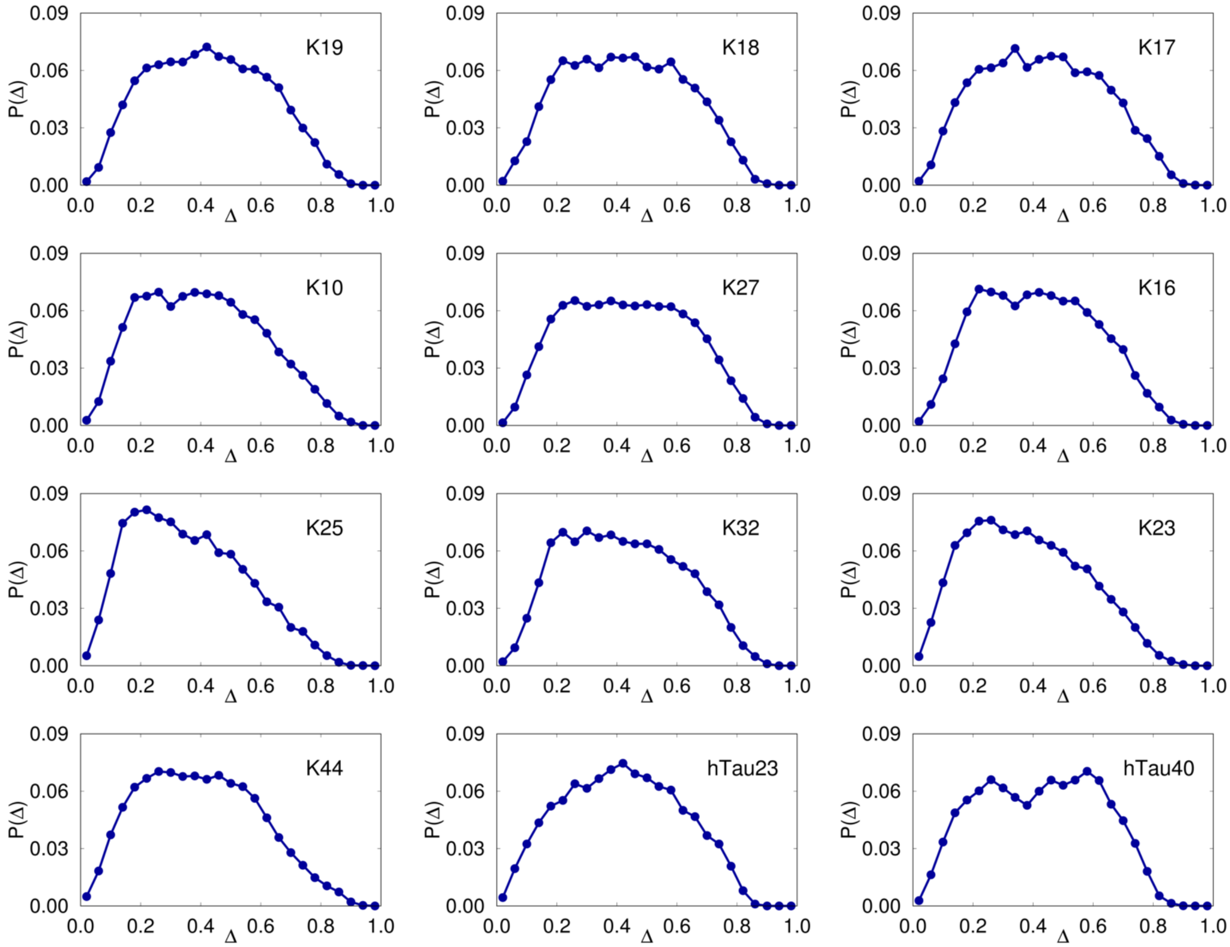

Figure S7: Same as Figure S6, except the results are for the Tau protein constructs.

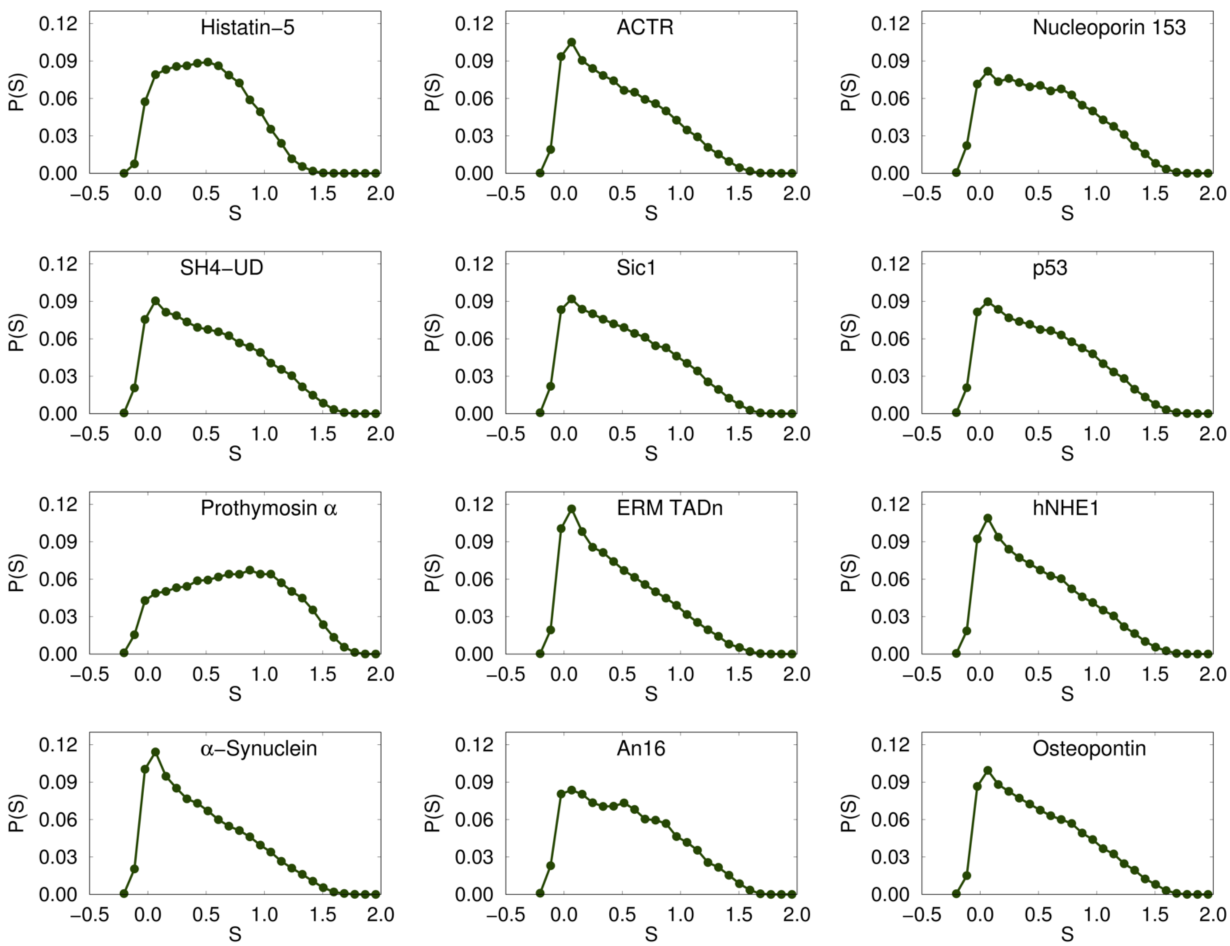

Figure S8: Distributions of the shape parameter S (eq 5 in the main text), for the twelve non-Tau IDP sequences.

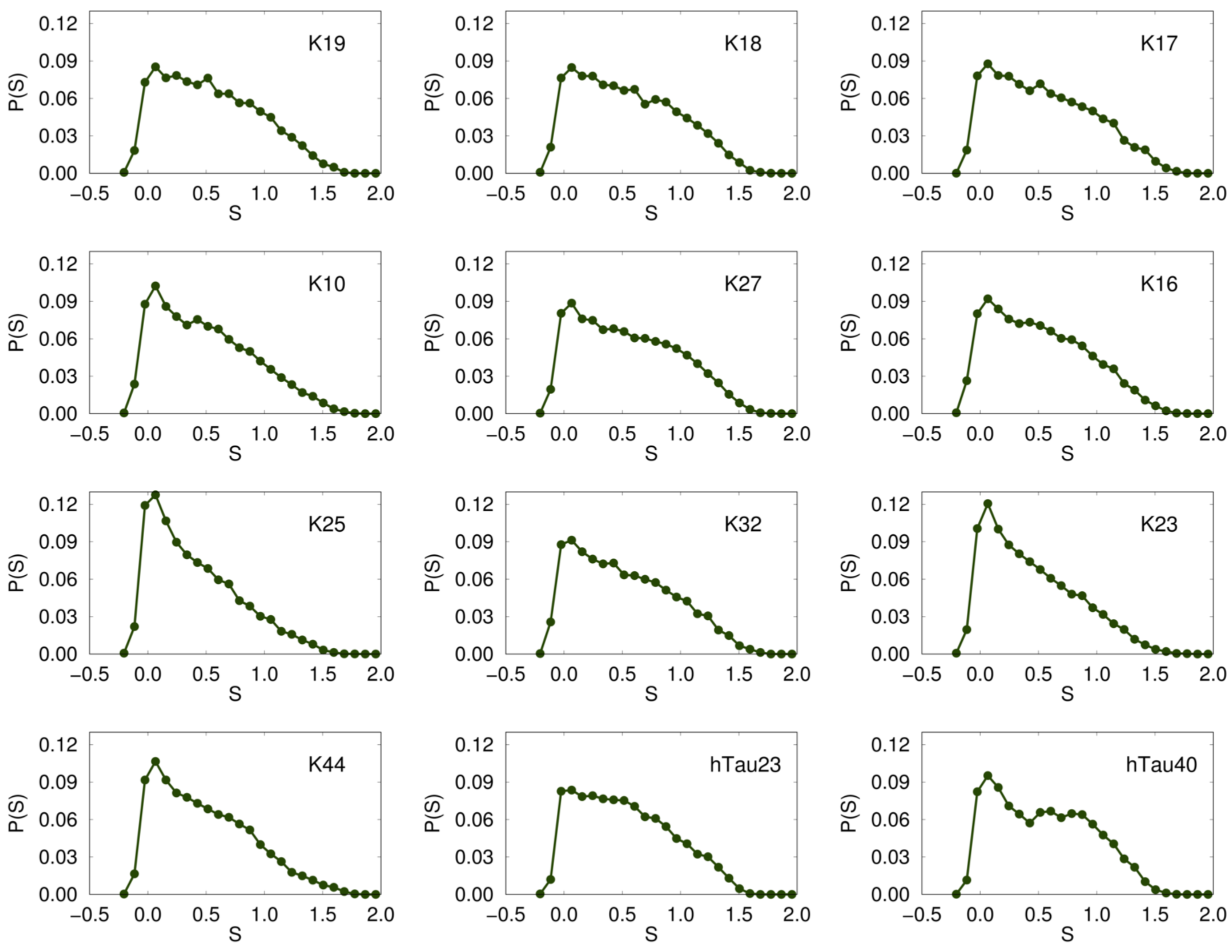

Figure S9: Same as Figure S8, except the results are for the Tau protein constructs.

Distribution of Radius of Gyration, R_g

In Figures S10 and S11, we show the R_g distributions for all the IDP sequences, which can be computed only in simulations. From experiments, one can only obtain the distance distribution functions by inverse Fourier transform of $I(q)$s. Nevertheless, these results are useful in assessing the extent of conformational fluctuations, and for benchmarking our simulations against other force fields.

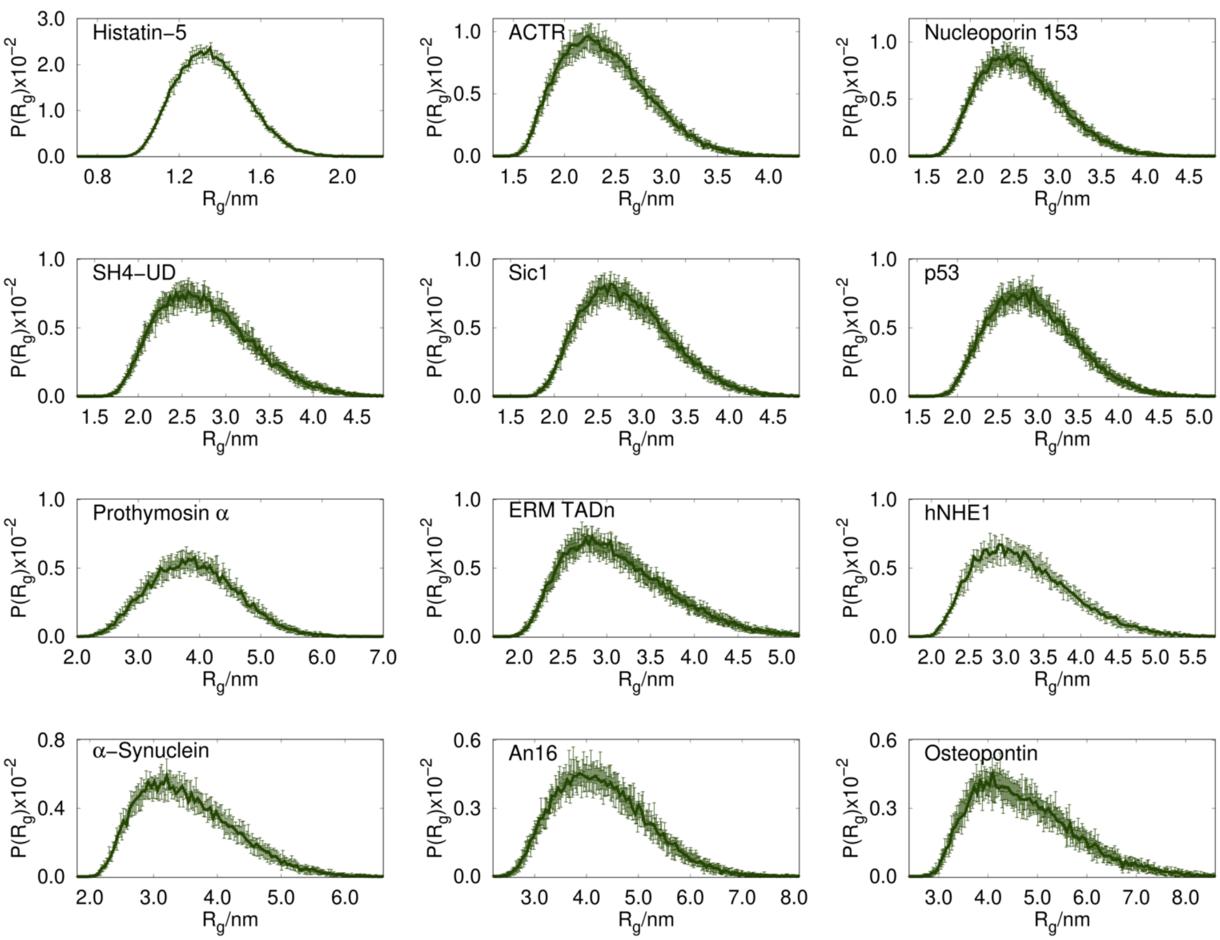

Figure S10: Distribution of R_g for all the non-Tau IDP sequences listed in Table S3.

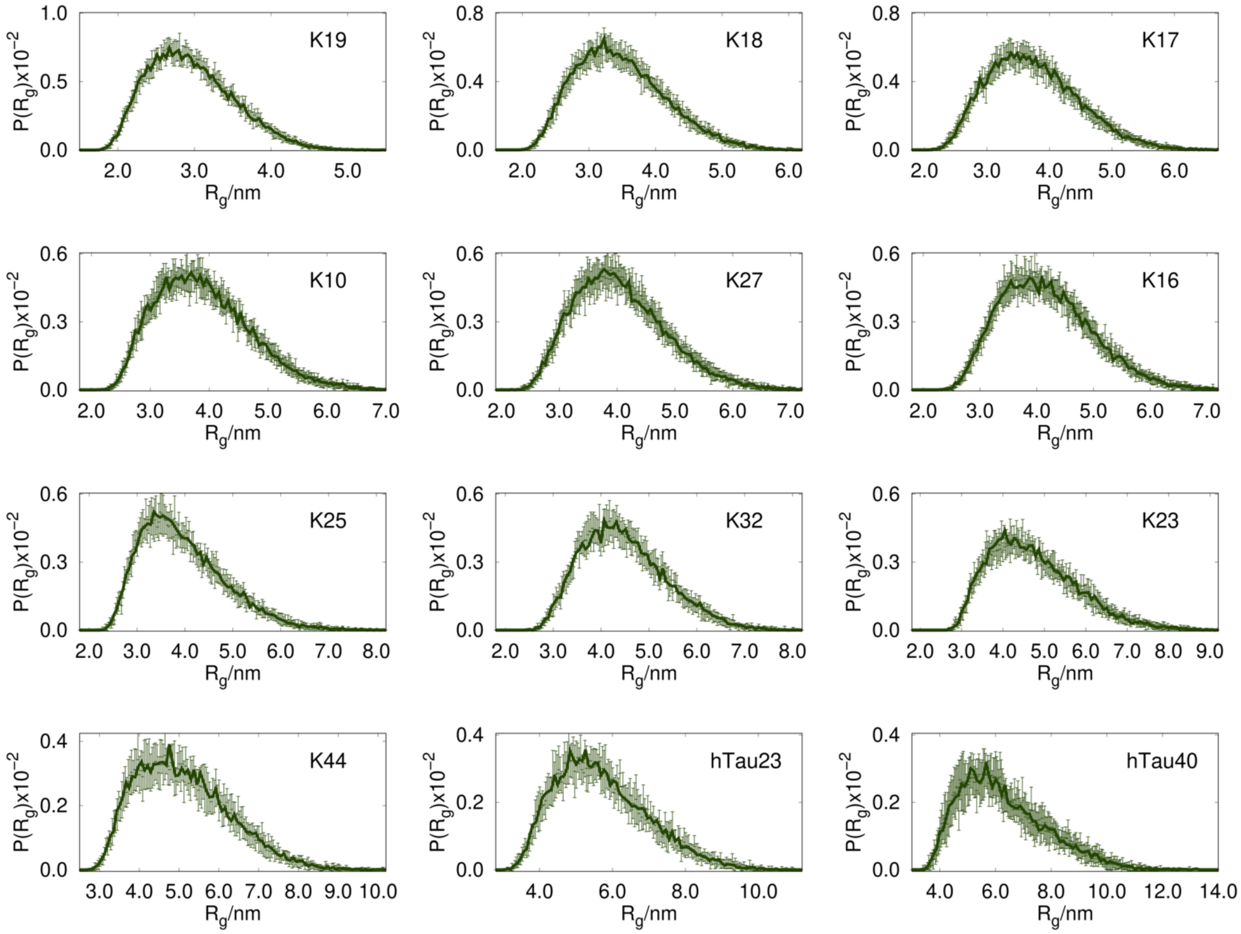

Figure S11: Distribution of R_g for all the Tau IDP sequences listed in Table S3.

End-end (R_{ee}) distance distributions, $P(R_{ee})$s.

In Figures S12 and S13, we show the R_{ee} distributions for all the IDP sequences, and compare them with rigorous theoretical curves corresponding to a random coil, and the Gaussian chain. The theoretical results for the random coil and Gaussian chain is given by[19] $P(x) = Cx^g e^{-\alpha x^\delta}$ for $N_T >> 1$, where $x = R_{ee}/\langle R_{ee}^2 \rangle^{\frac{1}{2}}$ and C is a normalization constant. The exponent $\delta = 1/(1-\nu)$ (with $\nu \approx 0.588$ for a random coil and $\nu = 0.5$ for a Gaussian chain) accounts for the decay of $P(x)$, for $x > 1$. The correlation hole exponent, g is ≈ 0.28 for a random coil, and represents the reduced probability of finding the end of the chains in contact in the presence of repulsive interactions. For a Gaussian chain, $g = 0$. The conditions $\int_0^\infty dx 4\pi x^2 P(x) = 1$ and $\int_0^\infty dx x^2 4\pi x^2 P(x) = 1$ allow the determination of C and α.[19] For a Gaussian chain $C = (3/2\pi)^{3/2}$, and $\alpha = 1.5$, whereas $C \approx 0.278$, and $\alpha \approx 1.206$ are approximate results for a random coil. In addition, when $N_T >> 1$, the effects of side-chains on $P(R_{ee})$ is unimportant, whereas for finite N_T, as is the case for IDPs considered here, they can be significant.

Interestingly, for the Tau protein constructs, the $P(R_{ee})$ distributions in the range $99 \leq N_T \leq 174$ (first six panels in Figure S13) are well described by the $P(R_{ee})$ corresponding to the theoretical random coil-like behavior, while in the range $185 \leq N_T \leq 441$ the distributions coincide with the predictions based on the Gaussian chain (last four panels in Figure S13). This is an example where the apparent solvent quality changes upon increasing N_T.

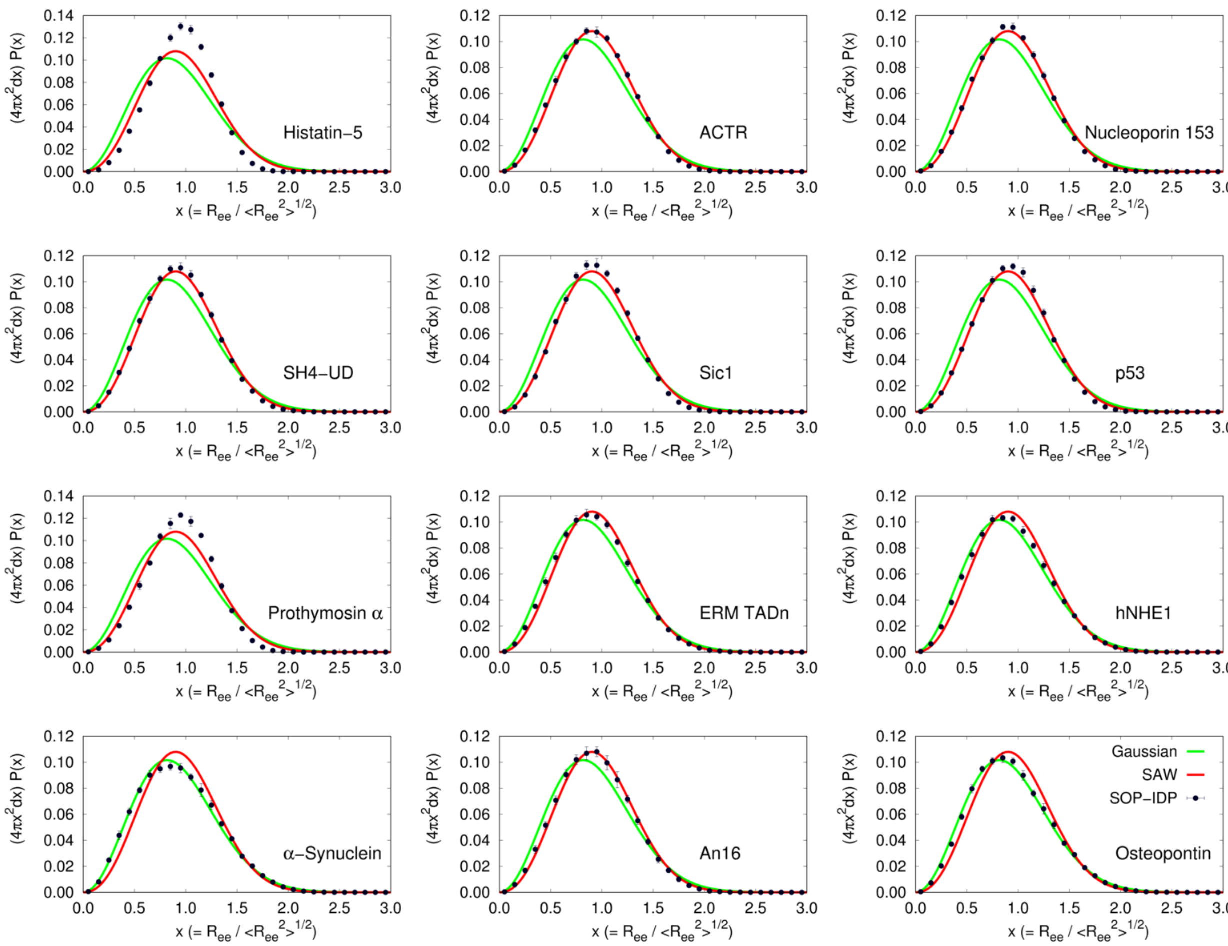

Figure S12: (black points) Plots for R_{ee} (scaled by $\langle R_{ee}^2 \rangle^{\frac{1}{2}}$ distance distributions for non-Tau IDP sequences. Theoretical distributions for random coil-like behavior and Gaussian chains are shown in red and green, respectively.

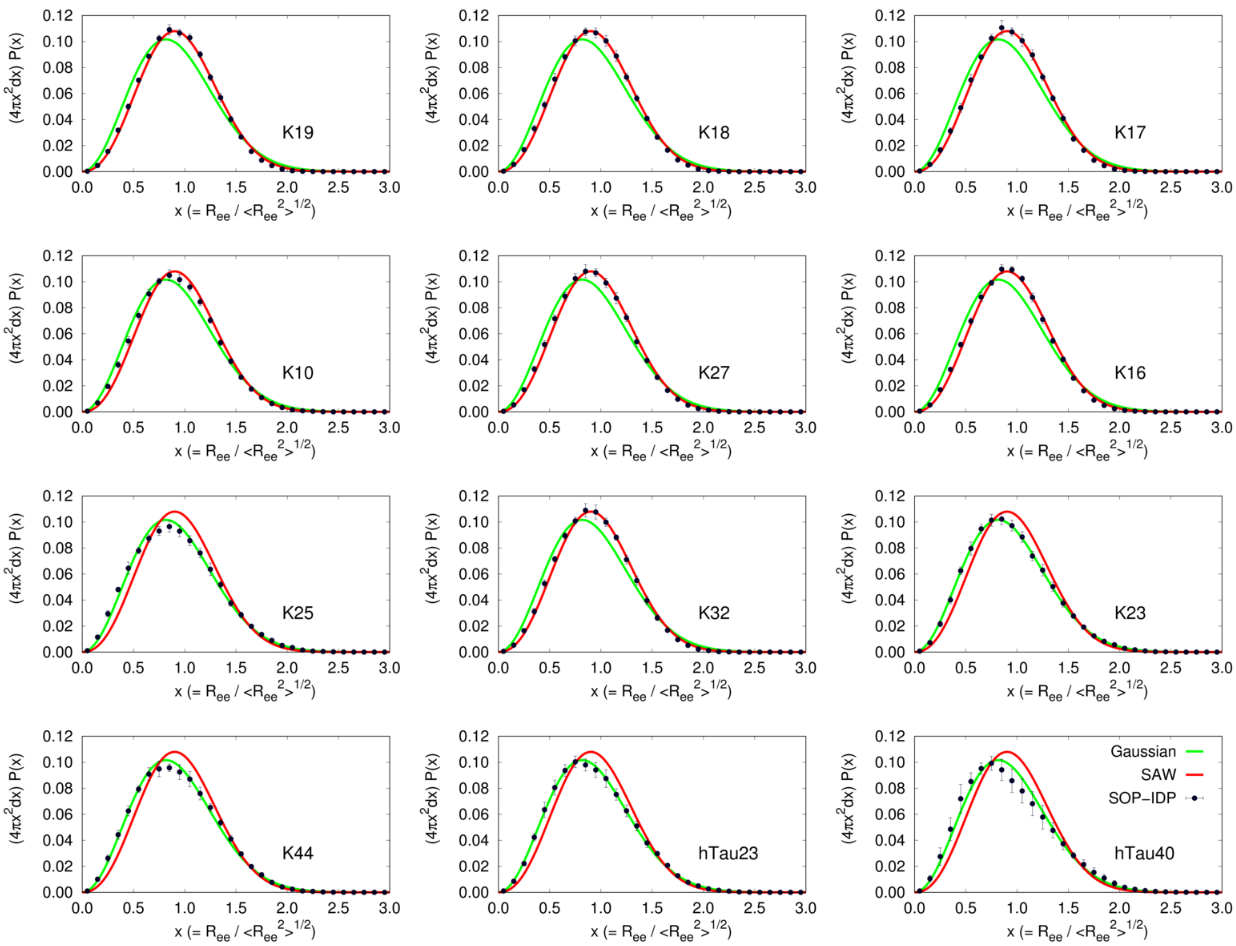

Figure S13: Same as Figure S12, except the results are for Tau protein constructs. Theoretical distributions for random coil-like behavior and Gaussian chains are shown in red and green, respectively.

Sequence compositional properties

Inspired by sequence variables that naturally arise in the theory of polyampholytes (PAs),[20,21] the biophysical properties of IDPs are often characterized on the basis of sequence compositional properties such as the fraction of positively / negatively charged residues (f_+ / f_-), net fraction of charged residues ($f_+ + f_-$), as well as quantities related to charge asymmetry, $\frac{(f_+ - f_-)^2}{(f_+ + f_-)}$.[22] Based on these variables it is natural to construct the plausible phases that a given IDP might adopt under ambient conditions. It is known from polyelectrolyte[23] and PA[20] theories that the phases could be readily altered by external conditions, such as changing the salt concentration. Clearly, the external conditions are not encoded in f_+ and f_-. Nevertheless, the PA type variables are useful in anticipating the states of IDPs. The PA-like properties for the IDP sequences are tabulated in Table S4 for all the simulated sequences. Although such quantities may be relevant in qualitative descriptions, their use in inferring conformational properties of specific IDP sequences is very limited. There is no obvious distinction in the compositional properties among the diverse set of IDPs studied in this work. While they span a wide range in their respective compositional properties, an understanding of their heterogeneous conformational ensembles necessarily requires explicit simulations, as discussed in detail in the main text.

Table S4: Sequence compositional properties of the studied IDP sequences.

IDP	f_+	f_-	$f_+ + f_-$	$(f_+ - f_-)^2/(f_+ + f_-)$
Histatin-5	0.29	0.08	0.38	0.116
ACTR	0.08	0.18	0.27	0.036
Nucleoporin 153	0.00	0.00	0.00	0.00
SH4-UD	0.14	0.07	0.21	0.024
Sic1	0.12	0.00	0.12	0.122
p53	0.02	0.18	0.20	0.127
Prothymosin α	0.09	0.49	0.58	0.273
ERM TADn	0.13	0.18	0.31	0.008
hNHE1	0.11	0.20	0.31	0.023
α-Synuclein	0.11	0.17	0.29	0.011
An16	0.03	0.0	0.03	0.032
Osteopontin	0.15	0.25	0.41	0.024
K19	0.20	0.09	0.29	0.042
K18	0.20	0.08	0.28	0.047
K17	0.20	0.07	0.27	0.064
K10	0.18	0.10	0.28	0.021
K27	0.21	0.08	0.29	0.060
K16	0.20	0.07	0.27	0.064
K25	0.17	0.13	0.30	0.005
K32	0.21	0.08	0.28	0.061
K23	0.16	0.13	0.29	0.004
K44	0.18	0.12	0.30	0.014
hTau23	0.17	0.12	0.29	0.011
hTau40	0.16	0.13	0.29	0.011

Contact maps reveal deviations from random coil-like behavior

As discussed in the main text, deviation from random coil-like behavior for certain IDPs, particularly the Tau constructs, can be gleaned from their difference contact maps. In Figure S14, we show that the WT Tau sequence (hTau40) contains a locally compact region, with enhanced contacts. This region is highlighted using a zoomed-in view of the contact map. The segment with locally enhanced contacts is also present in K25 (see Figure 8 in the main manuscript for a zoomed-in view, and below for the full range), and the K23, K44, and hTau23 constructs (see below).

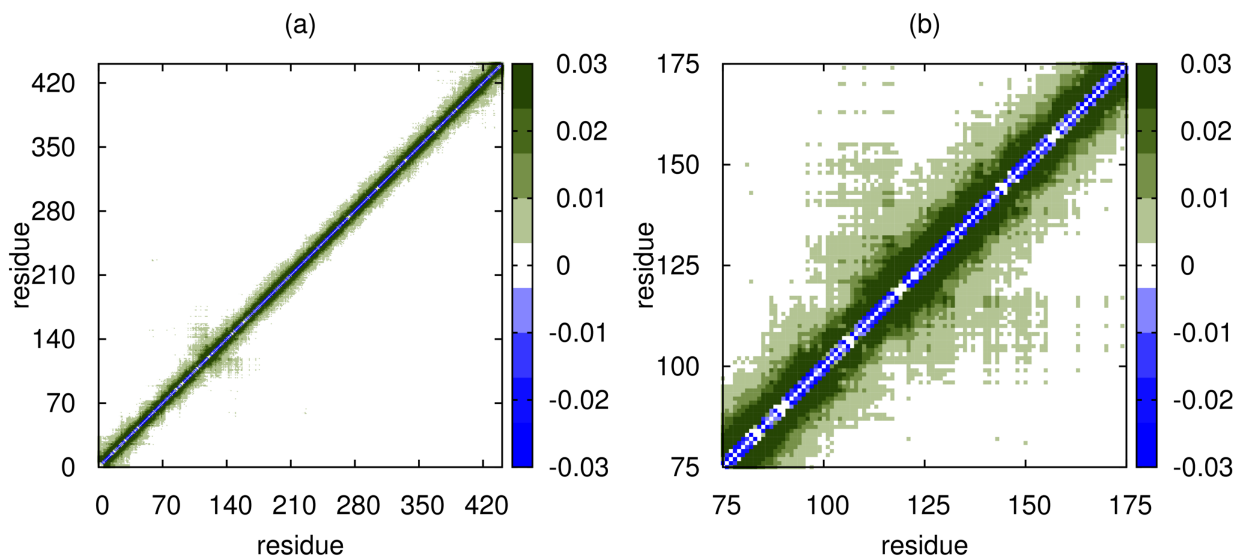

Figure S14: Difference contact map (defined in eq 7 in the main text) between the one obtained using SOP-IDP simulations and the corresponding contact map for a Flory random coil for (a) the full length Tau protein (hTau40). (b) Zoomed-in view, highlighting the region with enhanced contacts.

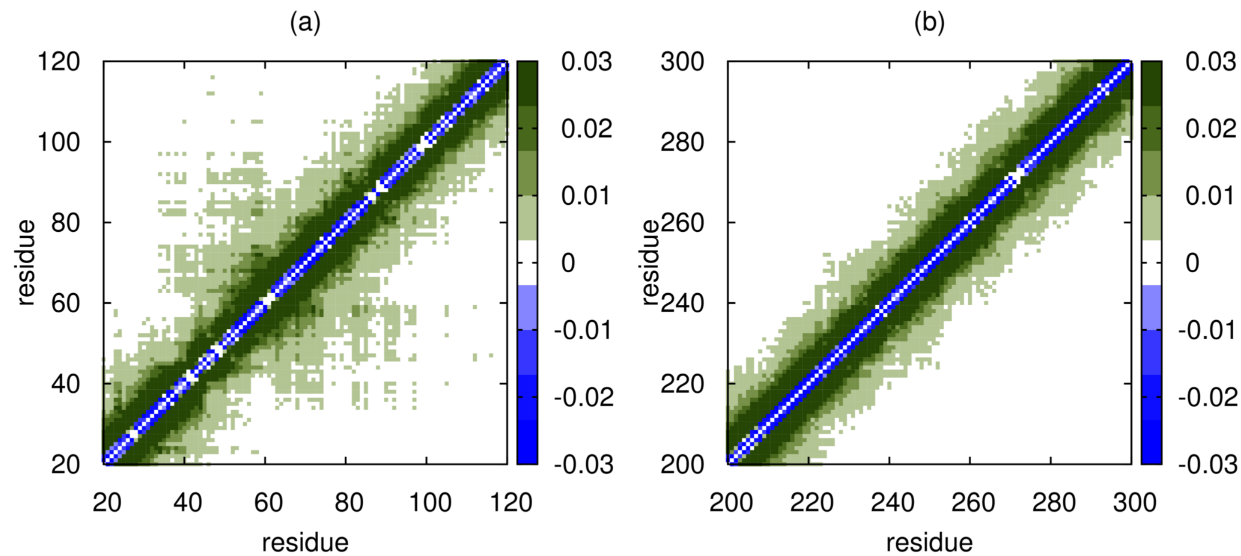

Figure S15: Same as Figure S14 except this is for the hTau23 sequence (a). Zoomed-in view for residues 200-300 of hTau40, shown for comparison (b).

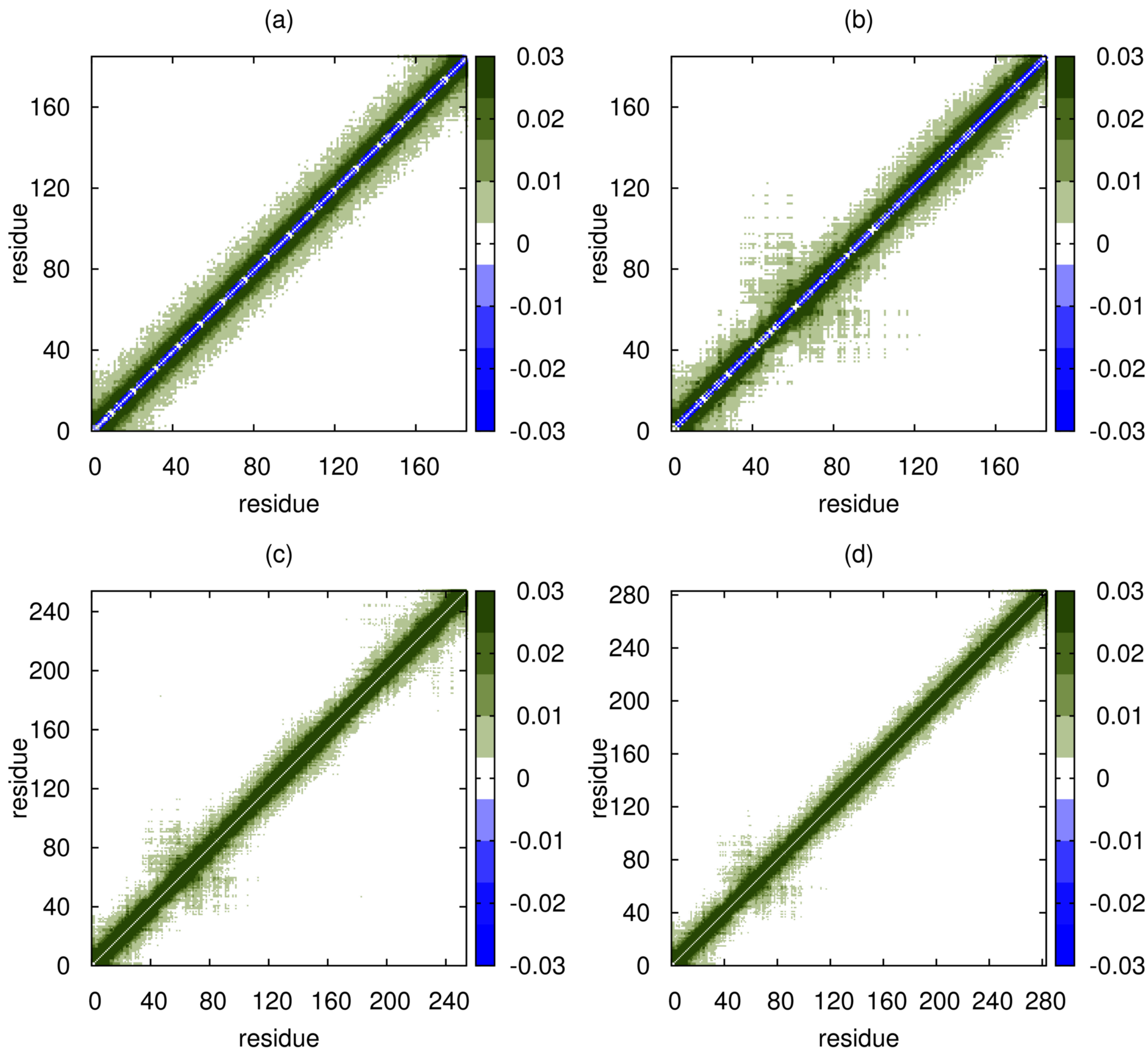

Figure S16: Difference contact maps (eq 7 in the main text) for (a) An16, (b) K25, (c) K23 and (d) K44, showing the full range in each case.

Simulated sequences for the IDPs in the FASTA representation

We use one letter codes for the amino acids. In cases where the sequences were not explicitly provided in the experimental references, the sequences were obtained from DisProt[24] database. As such, some sequences differ by ≤ 4 residues from the experimental sequence lengths. In such cases, the R_g values were scaled up/down following the appropriate N_T^{ν} scaling.

- Histatin-5

  DSHAKRHHGYKRKFHEKHHSHRGY

- ACTR

  GTQNRPLLRNSLDDLVGPPSNLEGQSDERALLDQLHTLLSNTDATGLEEIDRALGIPELVNQGQA
  LEPKQD

- Nucleoporin 153

  GCPSASPAFGANQTPTFGQSQGASQPNPPGFGSISSSTALFPTGSQPAPPTFGTVSSSSQPPVFG
  QQPSQSAFGSGTTPNA

- SH4-UD

  MGSNKSKPKDASQRRRSLEPAENVHGAGGGAFPASQTPSKPASADGHRGPSAAFAPAAAEPKLFG
  GFNSSDTVTSPQRAGPLAGG

- Sic1

    MTPSTPPRSRGTRYLAQPSGNTSSSALMQGQKTPQKPSQNLVPVTPSTTKSFKNAPLLAPPNSNM
    GMTSPFNGLTSPQRSPFPKSSVKRT

- p53

    MEEPQSDPSVEPPLSQETFSDLWKLLPENNVLSPLPSQAMDDLMLSPDDIEQWFTEDPGPDEAPR
    MPEAAPPVAPAPAAPTPAAPAPAPSWPL

- Prothymosin α

    MSDAAVDTSSEITTKDLKEKKEVVEEAENGRDAPANGNAENEENGEQEADNEVDEEEEEGGEEEE
    EEEEGDGEEEDGDEDEEAESATGKRAAEDDEDDDVDTKKQKTDEDD

- ERM TADn

    MDGFYDQQVPFMVPGKSRSEECRGRPVIDRKRKFLDTDLAHDSEELFQDLSQLQEAWLAEAQVPD
    DEQFVPDFQSDNLVLHAPPPTKIKRELHSPSSELSSCSHEQALGANYGEKCLYNYCA

- hNHE1

    MVPAHKLDSPTMSRARIGSDPLAYEPKEDLPVITIDPASPQSPESVDLVNEELKGKVLGLSRDPAK
    VAEEDEDDDGGIMMRSKETSSPGTDDVFTPAPSDSPSSQRIQRCLSDPGPHPEPGEGEPFFPKGQ

- α-Synuclein

MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVHGVATVAEKTKEQVTN
VGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKDQLGKNEEGAPQEGILEDMPVDPDNEAYEMPSE
EGYQDYEPEA

- An16

MHHHHHHPGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGA
PAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYG
APAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYGAPAQTPSSQYV

- Osteopontin

MRIAVICFCLLGITCAIPVKQADSGSSEEKQTLPSKSNESHDHMDDMDDEDDDDHVDSQDSIDSN
DSDDVDDTDDSHQSDESHHSDESDELVTDFPTDLPATEVFTPVVPTVDTYDGRGDSVVYGLRSKS
KKFRRPDIQYPDATDEDITSHMESEELNGAYKAIPVAQDLNAPSDWDSRGKDSYETSQLDDQSAE
THSHKQSRLYKRKANDESNEHSDVIDSQELSKVSREFHSHEFHSHEDMLVVDPKSKEEDKHLKFR
ISHELDSASSEVN

- K19

MQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVE
VKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIE

- K18

MQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIINKKLDLSNVQSKCGSKDNIKHVPGGGSVQ
IVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIE

- K17

  `MSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKSKIG`
  `STENLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSL`
  `DNITHVPGGGNKKIE`

- K10

  `MQTAPVPMPDLKNVKSKIGSTENLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVE`
  `VKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIETHKLTFRENAKAKTDHGAEIVYKSPVVSGDT`
  `SPRHLSNVSSTGSIDMVDSPQLATLADEVSASLAKQGL`

- K27

  `MSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKSKIG`
  `STENLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSL`
  `DNITHVPGGGNKKIETHKLTFRENAKAKTDHGAEIVY`

- K16

  `MSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKSKIG`
  `STENLKHQPGGGKVQIINKKLDLSNVQSKCGSKDNIKHVPGGGSVQIVYKPVDLSKVTSKCGSLG`
  `NIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIE`

- K25

MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEAAGHVT
QARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAAPPGQKGQANATRIPAKTPPAPKTPPSSGEP
PKSGDRSGYSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRL

- K32

MSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKSKIGS
TENLKHQPGGGKVQIINKKLDLSNVQSKCGSKDNIKHVPGGGSVQIVYKPVDLSKVTSKCGSLGNI
HHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIETHKLTFRENAKAKTDHGAEIVY

- K23

MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEAAGHVT
QARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAAPPGQKGQANATRIPAKTPPAPKTPPSSGEP
PKSGDRSGYSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLTHKLTFRENA
KAKTDHGAEIVYKSPVVSGDTSPRHLSNVSSTGSIDMVDSPQLATLADEVSASLAKQGL

- K44

MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEAAGHVT
QARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAAPPGQKGQANATRIPAKTPPAPKTPPSSGEP
PKSGDRSGYSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDL
KNVKSKIGSTENLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDFKDR
VQSKIGSLDNITHVPGGGNKKIE

- hTau23

MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEAAGHVT

QARMVSKSKDGTGSDDKKAKGADGKTKIATPRGAAPPGQKGQANATRIPAKTPPAPKTPPSSGEP

PKSGDRSGYSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDL

KNVKSKIGSTENLKHQPGGGKVQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQVEVKSEKLDFKDR

VQSKIGSLDNITHVPGGGNKKIETHKLTFRENAKAKTDHGAEIVYKSPVVSGDTSPRHLSNVSST

GSIDMVDSPQLATLADEVSASLAKQGL

- hTau40

MAEPRQEFEVMEDHAGTYGLGDRKDQGGYTMHQDQEGDTDAGLKESPLQTPTEDGSEEPGSETSD

AKSTPTAEDVTAPLVDEGAPGKQAAAQPHTEIPEGTTAEEAGIGDTPSLEDEAAGHVTQARMVSK

SKDGTGSDDKKAKGADGKTKIATPRGAAPPGQKGQANATRIPAKTPPAPKTPPSSGEPPKSGDRS

GYSSPGSPGTPGSRSRTPSLPTPPTREPKKVAVVRTPPKSPSSAKSRLQTAPVPMPDLKNVKSKI

GSTENLKHQPGGGKVQIINKKLDLSNVQSKCGSKDNIKHVPGGGSVQIVYKPVDLSKVTSKCGSL

GNIHHKPGGGQVEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIETHKLTFRENAKAKTDHGA

EIVYKSPVVSGDTSPRHLSNVSSTGSIDMVDSPQLATLADEVSASLAKQGL

References

(1) Liu, Z.; Reddy, G.; O'Brien, E. P.; Thirumalai, D. Collapse kinetics and chevron plots from simulations of denaturant-dependent folding of globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 7787–7792.

(2) Reddy, G.; Thirumalai, D. Dissecting ubiquitin folding using the self-organized polymer model. *J. Phys. Chem. B* **2015**, *119*, 11358–11370.

(3) Klimov, D. K.; Thirumalai, D. Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins* **2001**, *43*, 465–475.

(4) Best, R. B.; Hummer, G.; Eaton, W. A. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 17874–17879.

(5) Cragnell, C.; Durand, D.; Cabane, B.; Skepö, M. Coarse-grained modeling of the intrinsically disordered protein Histatin 5 in solution: Monte Carlo simulations in combination with SAXS. *Proteins: Struct., Funct., Bioinf.* **2016**, *84*, 777–791.

(6) Kjaergaard, M.; Nørholm, A.-B.; Hendus-Altenburger, R.; Pedersen, S. F.; Poulsen, F. M.; Kragelund, B. B. Temperature-dependent structural changes in intrinsically disordered proteins: Formation of α–helices or loss of polyproline II? *Prot. Sci.* **2010**, *19*, 1555–1564.

(7) Henriques, J.; Skepö, 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.* **2016**, *12*, 3407–3415.

(8) Best, R. B.; Zheng, W.; Mittal, J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J. Chem. Theory Comput.* **2014**, *10*, 5113–5124.

(9) O'Brien, E. P.; Ziv, G.; Haran, G.; Brooks, B. R.; Thirumalai, D. Effects of denaturants and osmolytes on proteins are accurately predicted by the molecular transfer model. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 13403–13408.

(10) Liu, Z.; Reddy, G.; Thirumalai, D. Theory of the molecular transfer model for proteins with applications to the folding of the src-SH3 Domain. *J. Phys. Chem. B* **2012**, *116*, 6707–6716.

(11) Liu, B.; Chia, D.; Csizmok, V.; Farber, P.; Forman-Kay, J. D.; Gradinaru, C. C. The effect of intrachain electrostatic repulsion on conformational disorder and dynamics of the Sic1 protein. *J. Phys. Chem. B* **2014**, *118*, 4088–4097.

(12) Uversky, V. N.; Gillespie, J. R.; MIllett, I. S.; ; Khodyakova, A. V.; Vasiliev, A. M.; Chernovskaya, T. V.; Vasilenko, R. N.; Kozlovskaya, G. D.; Dolgikh, D. A. et al. Natively unfolded human Prothymosin-α adopts partially folded collapsed conformation at acidic pH. *Biochemistry* **1996**, *38*, 15009–15016.

(13) Paleologou, K. E.; Schmid, A. W.; Rospigliosi, C. C.; Kim, H.-Y.; Lamberto, G. R.; Fredenburg, R. A.; Lansbury, P. T.; Fernandez, C. O.; Eliezer, D.; Zweckstetter, M. et al. Phosphorylation at Ser–129 but not the phosphomimics S129E/D inhibits the fibrillation of α–Synuclein. *J. Biol. Chem.* **2008**, *283*, 16895–16905.

(14) Wu, H.-N.; Jiang, F.; Wu, Y.-D. Significantly improved protein folding thermodynamics using a dispersion-corrected water model and a new residue-specific force field. *J. Phys. Chem. Lett.* **2017**, *8*, 3199–3205.

(15) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubműller, H. Structural ensembles of intrinsically disordered proteins depend strongly on force field: a comparison to experiment. *J. Chem. Theory Comput.* **2015**, *11*, 5513–5524.

(16) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, 4758–4766.

(17) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J. Phys. Chem. B* **2015**, *119*, 5113–5123.

(18) Rohtagi, A. WebPlotDigitizer. `https://automeris.io/WebPlotDigitizer`.

(19) Rubinstein, M.; Colby, R. H. *Polymer physics*; Oxford University Press: New York, U. S. A., 2003.

(20) Higgs, P. G.; Joanny, J. Theory of polyampholyte solutions. *J. Chem. Phys.* **1991**, *94*, 1543–1554.

(21) Ha, B. Y.; Thirumalai, D. Persistence length of intrinsically stiff polyampholyte chains. *J. Phys. II* **1997**, *7*, 887–902.

(22) Das, R. K.; Ruff, K. M.; Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2015**, *32*, 102–112.

(23) Ha, B.-Y.; Thirumalai, D. Conformations of a polyelectrolyte chain. *Phys. Rev. A* **1992**, *46*, R3012–R3015.

(24) Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; *et. al.*, DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **2017**, *45*, D219–D227.