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

Abstract

Intrinsically disordered proteins (IDPs), which are present in large numbers in eukaryotic proteomes, lack well-defined three-dimensional structures, thus challenging the archetypal notion of structure-function relationships. Determining the ensemble of conformations that IDPs explore under physiological conditions is the first step in understanding their diverse cellular functions. Here, we quantitatively characterize the structural features of IDPs as a function of sequence and length using molecular simulations of a coarse-grained model. For diverse IDP sequences, with the number of residues (N_T) ranging from 24 to 441, our simulations not only reproduce the radii of gyration (R_g) obtained from experiments, with the maximum deviation being no more than $\approx$ 18%, but also predict the full scattering intensity profiles in quantitative agreement with Small Angle X-ray Scattering experiments. The R_g values are well-described by the standard Flory scaling law, $R_g = R_0 N_T^{\nu}$, with $\nu \approx 0.588$, making it tempting to assert

that IDPs behave as polymers in a good solvent. However, clustering analysis reveals a menagerie of structures explored by IDPs is diverse, with the extent of heterogeneity being highly sequence-dependent, even though ensemble-averaged properties, such as the dependence of R_g on chain length, may suggest synthetic polymer-like behavior in a good solvent. For example, we show that for the highly charged Prothymosin-α a substantial fraction of conformations is highly compact. Even if the sequence compositions are similar, as is the case for α-Synuclein and a truncated construct from the Tau protein, there are substantial differences in the conformational heterogeneity. Taken together, these observations imply that metrics based on net charge or related quantities alone cannot be used to anticipate the phases of IDP, either in isolation or in complex with partner IDPs or RNA. Our work sets the stage for probing the interactions of IDPs with each other, with folded protein domains, or with partner RNAs, which are critical for describing the structures of stress granules and biomolecular condensates with important cellular functions.

Introduction

The discovery that a large fraction of eukaryotic protein sequences are not ordered in isolation[1–7] has produced a paradigm shift in the commonly held view of structure-function relationship. Such sequences, referred to as Intrinsically Disordered Proteins (IDPs), play key functional roles in diverse cellular processes, such as signal transduction,[2,8] and vesicular transport,[9] and are also implicated in neurodegenerative disorders and other diseases.[3,10] Experimental realizations of folding coupled to binding partners, and the more recent discovery of intracellular liquid-liquid phase separation[11] have further resulted in a concerted effort in describing the intramolecular and intermolecular interactions involving IDPs.[12–17] Recent estimates indicate that between (30-40)% of eukaryotic proteomes are either intrinsically disordered, or contain intrinsically disordered regions in otherwise folded proteins.[17,18] Despite the importance of IDPs, characterization of the sequence-dependent conformational proper-

ties of IDPs in isolation, as well as their phase behavior remains largely qualitative.[14,16,19] Biologically relevant forms of IDPs do not fold into unique tertiary structures, but explore a large number of distinct conformational states. In this sense, they are like synthetic polymers, whose conformations can only be characterized in statistical terms, such as the distribution of the radius of gyration, R_g, from which its dependence on the number of residues, N_T, can be calculated.[17,20] Two decades of experiments indicate that isolated IDPs typically behave as polymers in a good solvent, based on the scaling of R_g with N_T.[21–23] However, the conformational ensembles of IDPs, which are instrumental in modulating their *in vivo* functionalities, strongly depend on the precise sequence, in addition to other external conditions, such as pH, temperature, and salt concentration.[24,25] Therefore, scaling laws alone cannot provide a faithful description of the physico-chemical properties of IDPs. The complex sequence dependence poses a great challenge to purely theoretical approaches,[26,27] and raises the need for computational models that can quantitatively account for the experimental data.

Small-angle X-ray scattering (SAXS) and single molecule Förster resonance energy transfer (smFRET) have found the greatest applicability in the study of IDP conformational properties.[21,28] In addition, Nuclear Magnetic Resonance (NMR) and fluorescence correlation spectroscopy (FCS)[24,28–31] have been used to characterize the structural ensembles, and conformational dynamics of IDPs. SAXS experiments, which measure structure factors, could be used to calculate the global conformational properties of IDPs, such as R_g and shape, while smFRET experiments are used to probe distances between specific labeled residues along the polypeptide chain. With SAXS, the measured structure factor, $I(q)$, as function of the wave vector q, at small values of q is often used to estimate R_g using the Guinier approximation.[32] However, the sequence-specific conformational ensembles are not directly available from SAXS experiments, and usually techniques based on ensemble-optimization are invoked to map scattering profiles to representative structures.[33]

A complementary approach involves force-field based computer simulations to generate IDP ensembles, which embody the key experimental observables. It is tempting to exploit an

all-atom representation for IDPs and the surrounding environment in order to characterize their conformations. However, the well-documented inaccuracies of current all-atom force fields in describing unfolded states, and in particular the IDPs, leaves much to be desired.[34,35] Systematic benchmarking has revealed that current force-fields not only produce overly compact IDP conformations, but also strongly deviate from one another in terms of ensemble properties.[36,37] IDP-specific force fields introduced in recent times[35,38,39] incorporate *ad hoc* corrections to the original parametrization, and often suffer from a lack of transferability. A simpler description of the protein molecule based on systematic coarse-graining, often proves to be far more efficacious, and such approaches have been useful in a number of applications.[15,40,41] In this work, we introduce a robust and transferable self-organized polymer (SOP) coarse-grained model for IDPs (SOP-IDP), which not only recapitulates the wealth of experimental SAXS data, but also delineates the complex interplay between sequence, structure, and heterogeneity aspects that have been largely unexplored in much of the previous studies.

We show that for IDPs of varying lengths, and sequence composition, the R_g values are in accord with Flory's scaling law. The calculated R_g values deviate from those extracted from SAXS data by only 18% at best. The simulations quantitatively reproduce the measured SAXS profiles, thus allowing us to provide insights into the shape, and conformational fluctuations, which govern their functions. The ensembles of conformations of all the IDPs are heterogeneous, with sequence being a key determinant of the relative populations of different substates. Our findings using SOP-IDP simulations suggest that the phases of IDPs cannot be predicted solely based on sequence-compositional properties, but requires complete statistical characterization of the IDP ensembles.

Results

Simulated and experimental SAXS profiles are in near quantitative agreement: We used the SOP-IDP model (see Methods and SI for details) to calculate a variety of measurable structural properties for different IDP sequences. Twelve of these, including the initial training set consisting of Histatin-5, ACTR and hNHE1, are sequences with lengths ranging from 24 to 273, and are unrelated in composition and biological functions.[42–53] The Kratky plots computed for these sequences along with comparisons to SAXS experiments (Figure 1) show that the agreement between our simulations and experiments is excellent over most of the qR_g values, with the low qR_g region representing the global structure of the IDPs. Remarkably, for each of the IDPs for which the experimental SAXS profiles could be obtained, we observe a near perfect agreement in this regime, signifying that the model accurately captures the global conformations. The agreement between the calculated and the experimental SAXS profiles is excellent for all the sequences explored well beyond the range of the validity of the Guinier approximation (q$R_g < 3$) indicating that the model not only accurately describes the global shapes of IDPs, but also likely captures the structural details on smaller length scales. In some cases, the agreement with experimental profiles deteriorates at $qR_g > 3$, perhaps owing to a combined effect of the increased experimental noise, and the coarse-grained nature of the model.

Considering that the sample set (as shown in Figure 1) includes IDP sequences of varying sequence composition, and charge densities, from completely uncharged (Nucleoporin 153) to highly charged (Prothymosin α, $\sim 58\%$ charged residues, net charge -44), we conclude that the proposed SOP-IDP model is not only robust to sequence variations, but also accurate and transferable for variations in charge densities. We would like to emphasize that it has been rather difficult to obtain such a remarkable conformity to experimental trends with existing IDP models.[15,53]

Wild-type Tau proteins and its variants: In notable SAXS experiments, Svergun and coworkers measured the structure factors of the Wild-type (WT) 441 residue human Tau

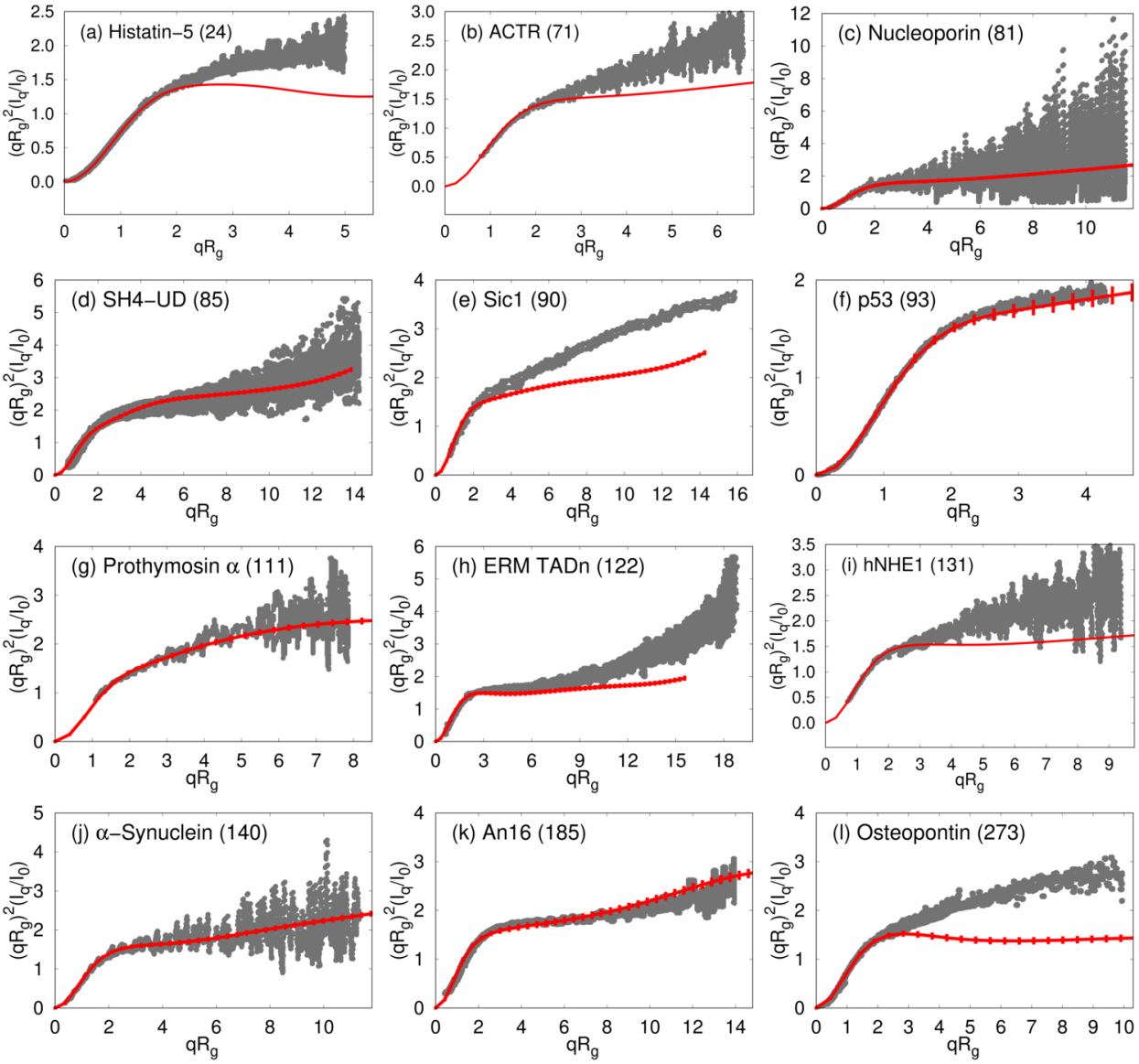


Figure 1: Comparison of SAXS profiles in the Kratky representation for 12 IDPs, labeled in the panels. The experimental profiles[42–52] are shown using gray dots, and those obtained from simulations are depicted using red curves. The numbers in parentheses in each figure is N_T, the number of residues in each IDP. Histatin-5 (a), ACTR (b) and hNHE1 (i) are part of the initial training set.

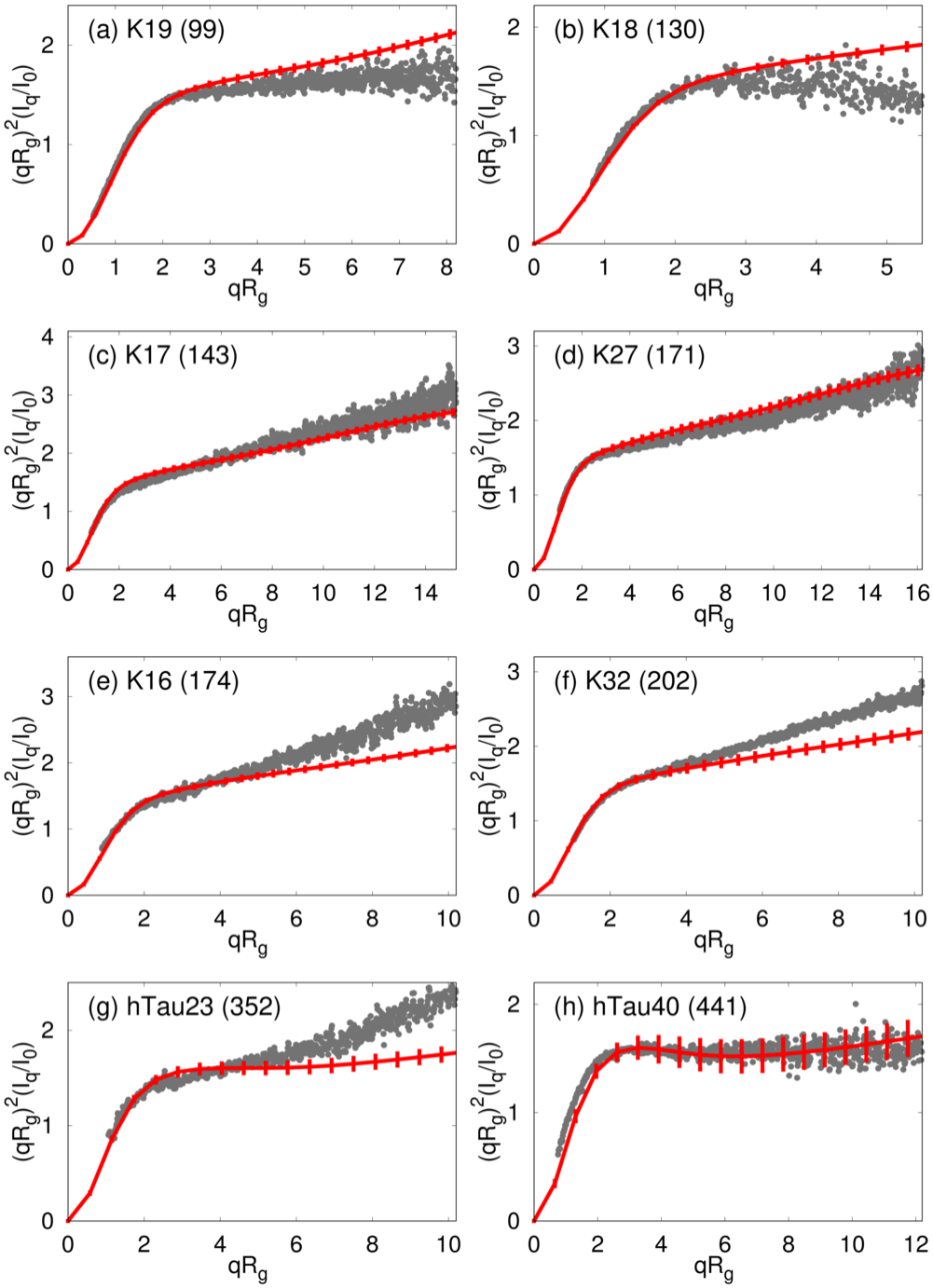


Figure 2: Comparison of SAXS profiles in the Kratky representation for eight Tau protein constructs. The lengths of the sequences are shown in parentheses in each figure. The color coding is the same as in Figure 1.

protein, as well as artificial sequence constructs generated through truncation and merger of segments.[54] We simulated eleven such artificial constructs (SI contains the sequences), in addition to the full length WT sequence using the SOP-IDP model, thus covering the full range of sequence lengths ($99 \leq N_T \leq 441$) considered in the experiments.[54] The Kratky plots for the Tau sequences are shown in Figure 2. The excellent agreement between the simulated and experimental profiles for IDPs of varying length and sequence composition clearly further establishes the transferability of the SOP-IDP model. It should be emphasized that we used only five (Histatin-5, ACTR, hNHE1, K32, and hTau40) of the twenty IDPs shown in Figures 1 and 2 to learn the parameters in the model (described in the SI). Of these, we used only the R_g values for K32 and hTau40 as a constraint to determine a single parameter, and not the SAXS profiles. Therefore, the quantitative agreement with experimental SAXS profiles is truly an emergent property of our model.

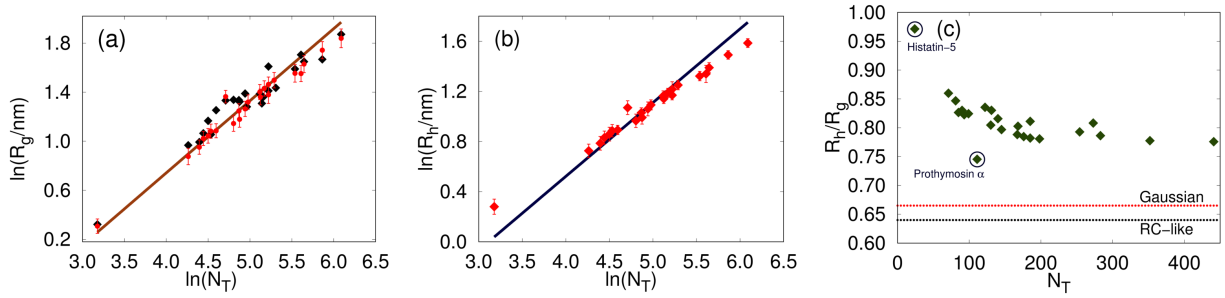


Figure 3: (a) Values of R_g from simulations (red circles) with comparison to the experimental estimates (black squares).The brown solid line is a fit to the power law: $R_g = R_g^0 N^\nu$, with $R_g^0 = 0.2\,\text{nm}$ and $\nu = 0.588$. (b) Hydrodynamic radii, R_hs, of the IDPs computed from the simulation trajectories (red symbols). The solid blue line is a power law fit: $R_h = R_h^0 N^\nu$, with $R_h^0 = 0.16\,\text{nm}$ and $\nu = 0.588$. Both R_g and R_h follow the Flory random coil predictions. (c) The ratio of the hydrodynamic radius to the radius of gyration R_h/R_g, obtained from simulations. The outliers, Histatin-5 and Prothymosin-α are marked with circles. The theoretical limits are marked as dotted lines. Note that there are substantial deviations from the theoretical predictions for ideal chain and polymer in a good solvent.

Dependence of radius of gyration and hydrodynamic radius on sequence length: Given the accurate reproduction of the SAXS profiles (Figure 1 and Figure 2), it is not surprising that the R_g values computed from simulations are in quantitative agreement with the experimental values. The scaling of R_g with sequence length (N_T), fitted using the

calculated or experimental R_g values, follows $R_g \approx 0.20 N_T^{\nu}$ nm, with $\nu \approx 0.588$ (Figure 3 (a)), which naively suggests that in the experimental conditions the IDPs might be in a good solvent. However, this is not the case (see below).

We also calculated the hydrodynamic radii, R_h, from the conformational ensemble using Eq. 4. The computed values of R_h are shown in Figure 3(b), and are tabulated in Table S2 of the SI. Though not equal in magnitude, both R_g and R_h (measurable from NMR for example) should show the same scaling with N_T with identical value of ν (Figures 3 (a,b)).[55] However, this scaling behavior does not imply that these IDPs behave as random coils.[32] Indeed, the R_h/R_g ratios from simulations deviate substantially from the established theoretical limits, which are 0.665 and 0.640, for an ideal chain, and a polymer in a good solvent, respectively (Figure 3(c)). As N_T increases beyond 300, R_h/R_g is insensitive to sequence length (Figure 3(c)), indicating that the observed deviations may not be due to finite size effects.

Distributions of radius of gyration (R_g) and end-end distances (R_{ee}): If the solution conditions used in experiments are good solvents for IDPs, as the R_g and R_h scaling with N_T imply (Figure 3), then it should be reflected in the distribution $P(R_{ee})$ of the the end-to-end distance (R_{ee}), which is rigorously known for a polymer in a good solvent as well as the ideal chain. Therefore, comparing the simulated distributions to the rigorous functional forms provides a stringent test of the solvent quality. The $P(R_{ee})$ distributions, with R_{ee} measured as the separation between the backbone atoms at the termini, are shown in Figures 4 (a,b) (see also supplementary Figures S7 and S8 in the SI). As a general rule, we find that the $P(R_{ee})$ for IDPs closely resemble the theoretical result for a random coil. However, for specific sequences the distributions are more skewed towards the ideal or Gaussian chain limit.[32,56] The deviations from the random coil predictions based on the theoretical predictions are likely due to non-local electrostatic interactions, and sequence composition (see below). Figure 4(a) shows that the $P(R_{ee})$ distribution for Prothymosin-α differs substantially from both the theoretical limits. The deviation can be attributed to the high fraction

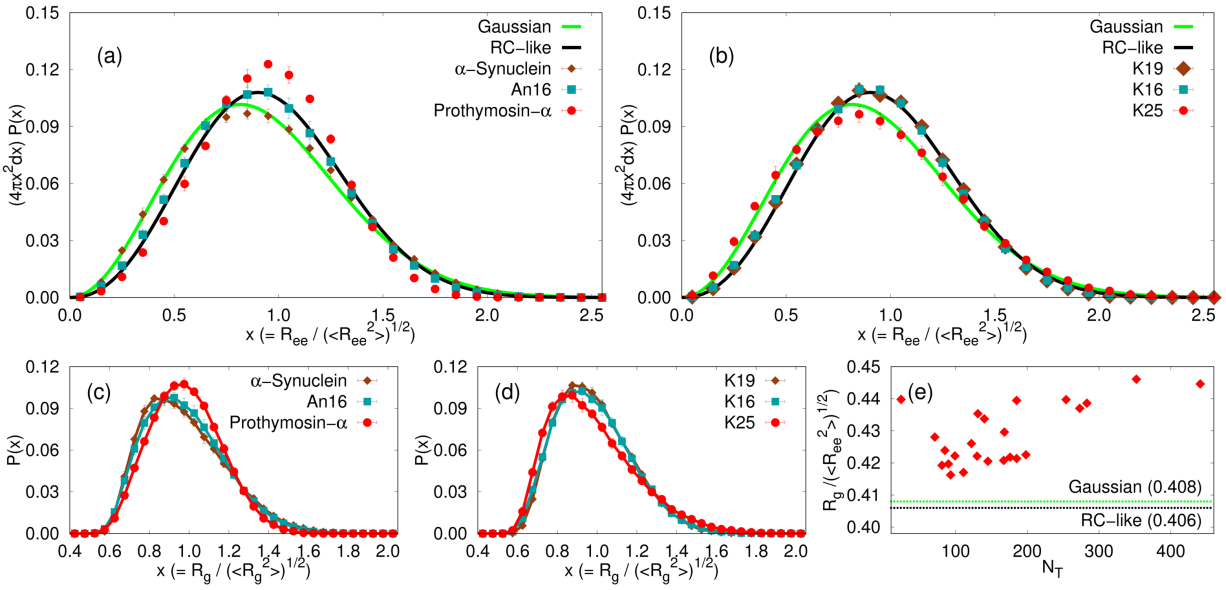


Figure 4: Distributions of the end to end distance (R_{ee}), scaled by $\langle R_{ee}^2 \rangle^{1/2}$. In order to compare with results for Gaussian chains (solid green line) and random coils (solid black line) we show $4\pi x^2 P(x)$, where $x = R_{ee}/\langle R_{ee}^2 \rangle^{1/2}$. Figures (a) and (b) show the distributions for representatives from the sets of non-Tau IDP sequences and Tau-sequence constructs, respectively. If a particular IDP behaves strictly as a random coil (Gaussian chain) in terms of end-to-end distance distribution, then the corresponding distribution should coincide exactly with the theoretical result in black (green). (c,d) Distributions of R_g for IDPs described in (a) and (b) respectively. (e) Ratio of R_g to root mean square R_{ee}, for all 24 IDPs described in the manuscript. The dashed horizontal lines in green and black show the ratio for a Gaussian chain and a random coil, respectively.

of charged residues ($\sim 58\%$).

Sequence matters: Interesting sequence-specific effects on the $P(R_{ee})$s are also observed for the Tau protein constructs, in spite of their common parent sequence. The $P(R_{ee})$ for the majority of the Tau protein constructs show excellent agreement with the theoretical $P(R_{ee})$ for random coil (see Figure 4 (b) for K19 and K16, and also Figure S8 in the SI), and deviate substantially from that for an ideal chain. Deviations from the random coil behavior are found for 5 out of the 12 Tau sequence constructs, namely K25 (Figure 4 (b)), K23, K44, hTau23 and the WT sequence, hTau40 (Figure S8 in the SI). The $P(R_{ee})$ for these sequences show increased propensities for smaller R_{ee} distances, as can be seen from Figure 4(b) for the K25 Tau construct. These sequence-specific deviations from average random coil-like behavior are also discernible in the distribution of R_g values, shown in Figures 4 (c) and (d). In Figure 4(e), the ratio R_g to root-mean-square R_{ee} is shown as a function of N_T. This ratio is 0.406 for a random coil, and 0.408 for a Gaussian chain.[56] Interestingly, even for those IDPs, whose $P(R_{ee})$ coincides with the theoretical predictions for random coils, the $R_g/\langle R_{ee}^2\rangle^{1/2}$ values deviate from the theoretical limit. This deviation becomes particularly pronounced as N_T increases, most notably for the longer Tau protein constructs.

The Tau sequences, which do not conform to standard polymer behavior, namely K25, K23, K44, hTau23, and hTau40, do not differ appreciably from the rest of the sequence constructs in terms of conventional sequence compositional parameters often used to rationalize their shapes (Table S3 in the SI). This apparent incongruity could be rationalized in terms of ensemble averaged contact maps obtained from simulations (see the Discussion section). We document below that deviation from the theoretical random coil limit, is in fact a direct manifestation of the sequence-dependent conformational heterogeneity.

Shape parameters are sequence dependent: The sequence-specific shape fluctuations of IDPs can be gleaned from the distributions of their shape parameters, Δ and S (see Eq. 5 in Methods). The calculated values of Δ, and S indicate that the conformations of IDPs can be described as prolate ellipsoids (Table S2 in the SI). In Figure 5, we show the distributions

for four IDP sequences, and the rest are included in the supporting information (Figures S3–S6).

Figure 5 (a) shows that Prothymosin-α has a high preference for extended conformations, while the structural ensembles for the K25 construct, and α-Synuclein are closer to being spherical. In contrast, the shape parameters for An-16 are homogeneously distributed, suggesting that elongated, and spherical conformations are equally probable. This systematic trend is also reflected in the distributions of S (Figure 5(b)), where the bias towards prolate conformations is maximal for Prothymosin-α, and minimal for K25. The large dispersions in $P(\Delta)$ and $P(S)$ makes the calculation of the mean values, listed in Table S2 of the SI, not meaningful (see SI for elaboration). For comparison, the average values of Δ and S for Gaussian chains are 0.52 and 0.87, and the corresponding estimates for an equivalent polymer in a good solvent are 0.55 and 0.91 respectively.[57] Overall, the shape fluctuations of the IDPs allude to a sequence-dependent heterogeneity of conformational ensembles. This aspect, which seems characteristic of the conformations of the IDPs explored here, is discussed in more detail below.

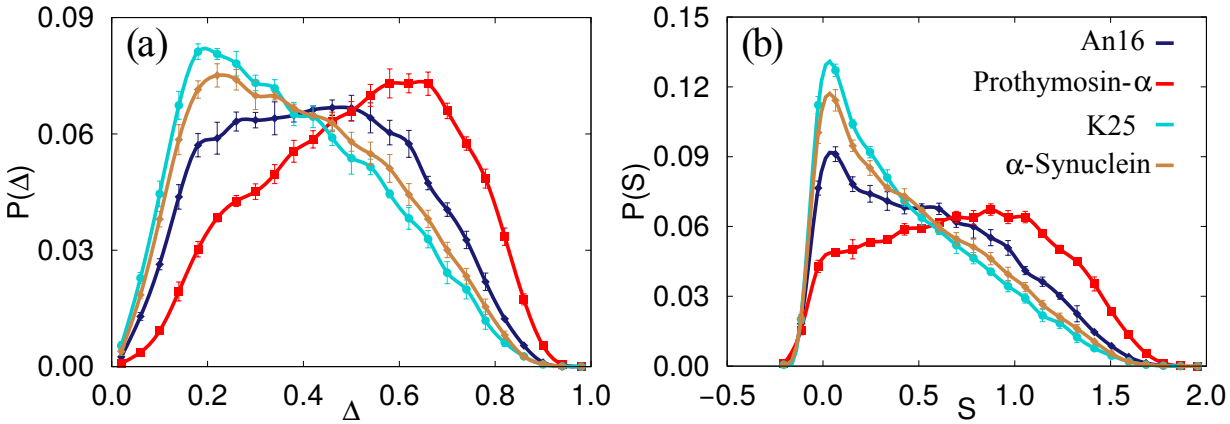


Figure 5: Comparison of the distributions of shape parameters (a) Δ and (b) S for An-16 (blue, diamonds) α-Synuclein (brown, diamonds), Prothymosin-α (red, squares) and K25 construct of the Tau protein (cyan, circles). Note that the average values of Δ and S for Gaussian chains are 0.52 and 0.87, and the corresponding estimates for an equivalent polymer in a good solvent are 0.55 and 0.91 respectively. The mean values of both Δ and S for the IDPs are drastically different from the expected theoretical values for standard polymer models.

IDP conformations are structurally heterogeneous: The distributions of R_g and R_{ee}, as well as their deviations from random coil-like behavior for certain IDPs, suggest that the equilibrium populations of different conformations populating the IDP ensemble depend on the sequence. To illustrate the importance of conformational fluctuations in determining the statistical properties of IDPs, we consider representative examples: Prothymosin-α, α-Synuclein, An-16, and the K25 construct from the family of Tau proteins.

To obtain insights into the structural ensembles, we performed hierarchical clustering of the IDP conformations. Hierarchical clustering not only provides a means to quantify the contrasting features of the structural ensemble, which are evident from the Δ and S distributions, but also aids in visualizing the extent of the underlying conformational heterogeneity. The results from the clustering analyses of α-Synuclein, and the K25 Tau construct are shown in Figure 6. The conformational ensembles corresponding to An16 and Prothymosin-α are depicted in Figure 7.

The appropriate number of clusters for each sequence was determined by evaluating the largest distance jumps in the corresponding dendrograms using the elbow method.[58] Figures 6 and 7 show that our clustering scheme is robust, and the various families are clearly demarcated on a two-dimensional projection of the conformational landscape onto the R_g, and R_{ee} coordinates.

α–Synuclein and K25 : The equilibrium ensemble of α-Synuclein (Figure 6) partitions into at least three clusters, and consists of a large population of relative compact structures (43.9%). Both semi-extended, and fully extended structures are populated to a lesser extent, with occupation probabilities of 27.5%, and 28.6%, respectively. As is evident from the dendrogram (Figure 6), the equilibrium ensemble of the K25 Tau protein construct is overwhelmingly dominated by relatively compact structures, with the overall contribution being 67.1% to the net population. This trend is also reflected in the two-dimensional conformational landscape, where the region with small R_g and R_{ee} values is associated with the highest density. The shape of the conformational landscapes for both α-Synuclein, and K25, as well as the

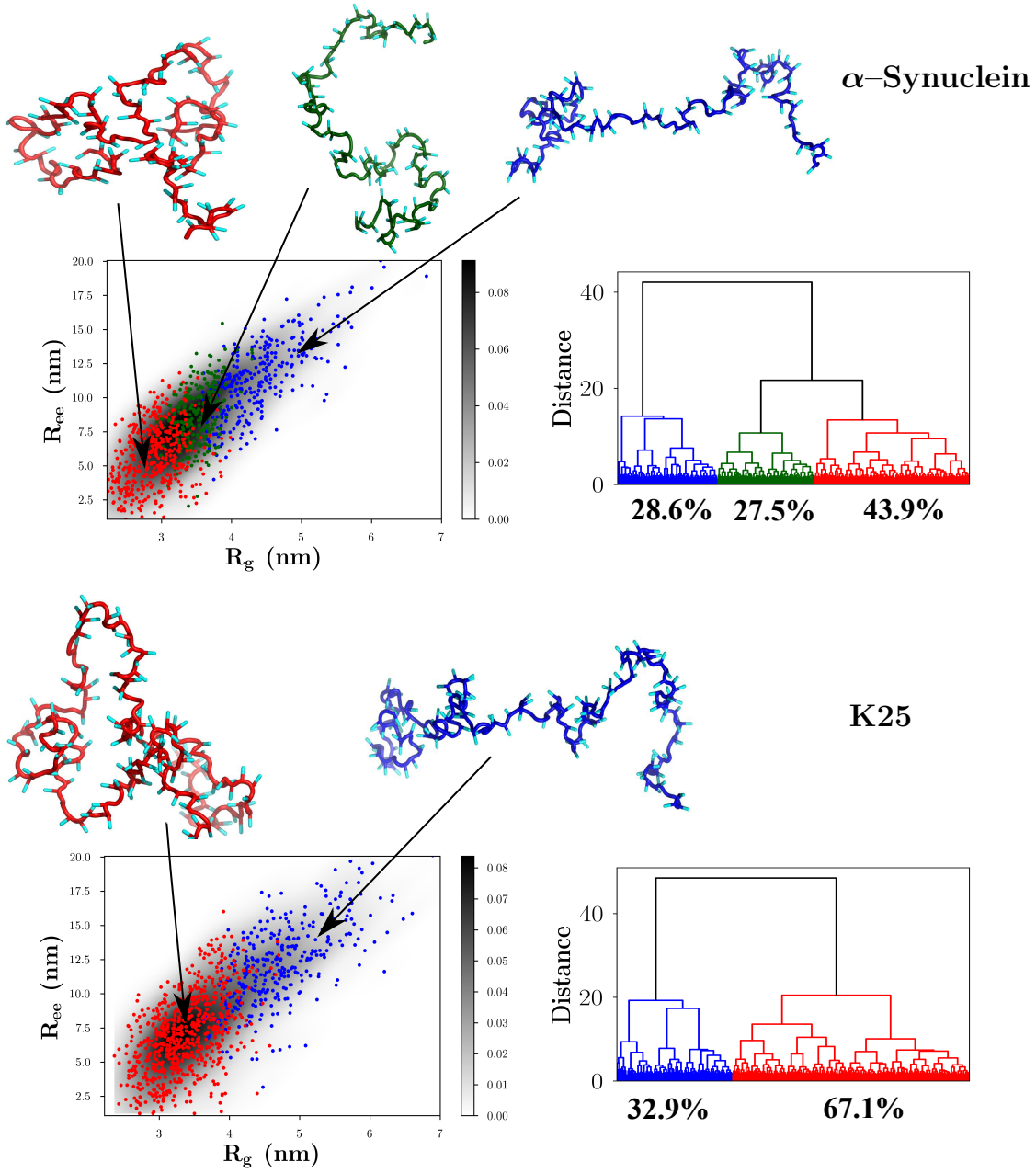


Figure 6: Hierarchical clustering of the IDP conformational ensembles using the Ward variance minimization algorithm. Upper panel: α-Synuclein; Lower panel: K25. The conformational landscapes (depicted as probability density plots) projected onto R_g and R_{ee} are shown on the left of each figure, and the corresponding dendrograms are shown on the right. Representative snapshots from each family, with their backbones rendered in different colors, are shown superposed on the conformational landscape. The same color-coding is used to classify clusters within the dendrograms, and the two-dimensional plots. The relative cluster populations are marked below the appropriate dendrogram branches.

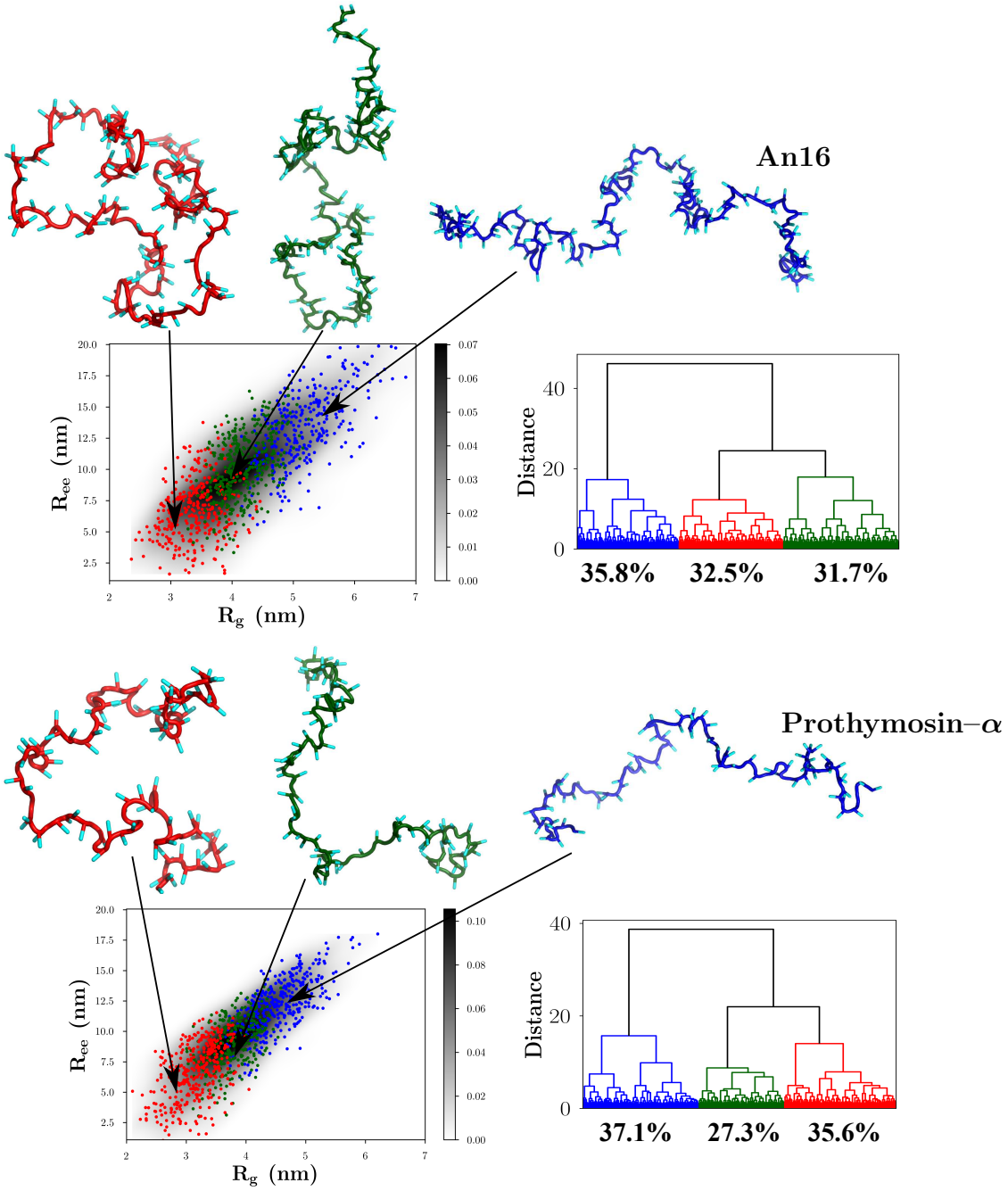


Figure 7: Upper panel: An16; Lower panel: Prothymosin-α. The conformational landscapes projected onto R_g and R_{ee} are shown on the left of each figure, and the corresponding dendrograms are shown on the right. Representative snapshots from each family, with their backbones rendered in different colors, are shown superposed on the conformational landscape. The same color-coding is used to classify clusters within the dendrograms, and the two-dimensional plots. The relative cluster populations are marked below the appropriate dendrogram branches.

relative cluster sizes are commensurate with the observed deviations from random coil-like behavior.

An16 and Prothymosin-α : The equilibrium ensemble of An16 (Figure 7) consists of approximately equal contributions from relatively compact, semi-extended, and extended structures, with the populations being 32.5%, 31.7% and 35.8%, respectively. The relative populations suggest that no specific conformation is particularly favored for An16, which is consistent with a random coil-like behavior.

For the Prothymosin-α sequence, the ensemble of conformations partitions into at least three clusters (Figure 7). Although naively we expect that, due to the high value of net charge, there would be an overwhelmingly large population of extended structures for Prothymosin-α, clustering of conformations shows otherwise. While conformations exhibiting high R_g and R_{ee} values contribute maximally to the equilibrium population (37.1%), relatively compact conformations constitute the second-largest cluster, and account for as much as $\approx$ 35.6% of the conformational ensemble. Configurations having intermediate values of R_g and R_{ee}, or the semi-extended structures have the lowest occupation probability (27.3%).

Overall, the systematic variations in the heterogeneity of sampled conformations for the different IDP sequences provide a structural explanation for the calculated distributions of R_{ee} and R_g, which are clearly masked if only their mean values are analyzed. It is indeed remarkable that although mean values of R_g and R_h adhere to Flory scaling laws for good solvents, there is considerable fine structure in the IDP ensembles, as the clustering analyses demonstrate. It is likely that the plasticity of the conformational ensembles makes it possible for IDPs to carry out diverse functional roles in a bewildering array of metabolic pathways.

Discussion

Enhancements in local contact probability leads to deviation from random coil behavior: A hallmark of polymers in good solvents (random coils or RCs) is that the

$P(R_{ee})$ must obey the theoretical distribution obtained using polymer theory (see SI). The statistical properties of the ensemble of conformations of RCs are purely determined only by chain entropy. Interestingly, we find that although the dependence of R_g on N_T for all the IDPs is consistent with the Flory scaling law (Figure 3) there are deviations in $P(R_{ee})$ from the expected theoretical predictions for RCs. For example, we observed such deviations for α-Synuclein, Prothymosin-α, as well as the Tau sequences K25, K23, K44, hTau23 and hTau40, but not for An16. The apparent conundrum can be understood using the dendrograms, which reveal clearly the sequence-dependent conformational heterogeneity. As a result the conformational ensembles are determined not only by entropy but also sequence-dependent energy, resulting in higher Boltzmann weight for certain classes of conformations relative to others.

To further illustrate the origin of deviation from the RC behavior, we calculated ensemble averaged contact maps from simulations using the SOP-IDP energy function, as well as the corresponding RC limit. We define the difference contact maps as,

$$\Delta C(i,j) \;=\; C_{IDP}(i,j) \;-\; C_{RC}(i,j) \tag{1}$$

where $C_{IDP}(i,j)$ is the contact probability between residues i,j of an IDP obtained from simulations, and $C_{RC}(i,j)$ is the same contact probability obtained by specifically neglecting the contributions due to the last four terms in Eq. 2 (see the Methods section). Thus, $\Delta C(i,j)$ is a measure of the increased or decreased probability of contact between two residues, compared to an equivalent RC. Both $C_{IDP}(i,j)$ and $C_{RC}(i,j)$ were computed using the coordinates of the side-chain beads, including glycine and alanine. We assume that a contact exists if the distance between two residues is $\leq 0.8\,\mathrm{nm}$.

The three contact maps shown in Figure 8 represent characteristically divergent scenarios. *An16:* The contact map for An16 (Figure 8 (a)) shows a uniform enhancement in contacts between residues proximate along the sequence, over the entire sequence length compared

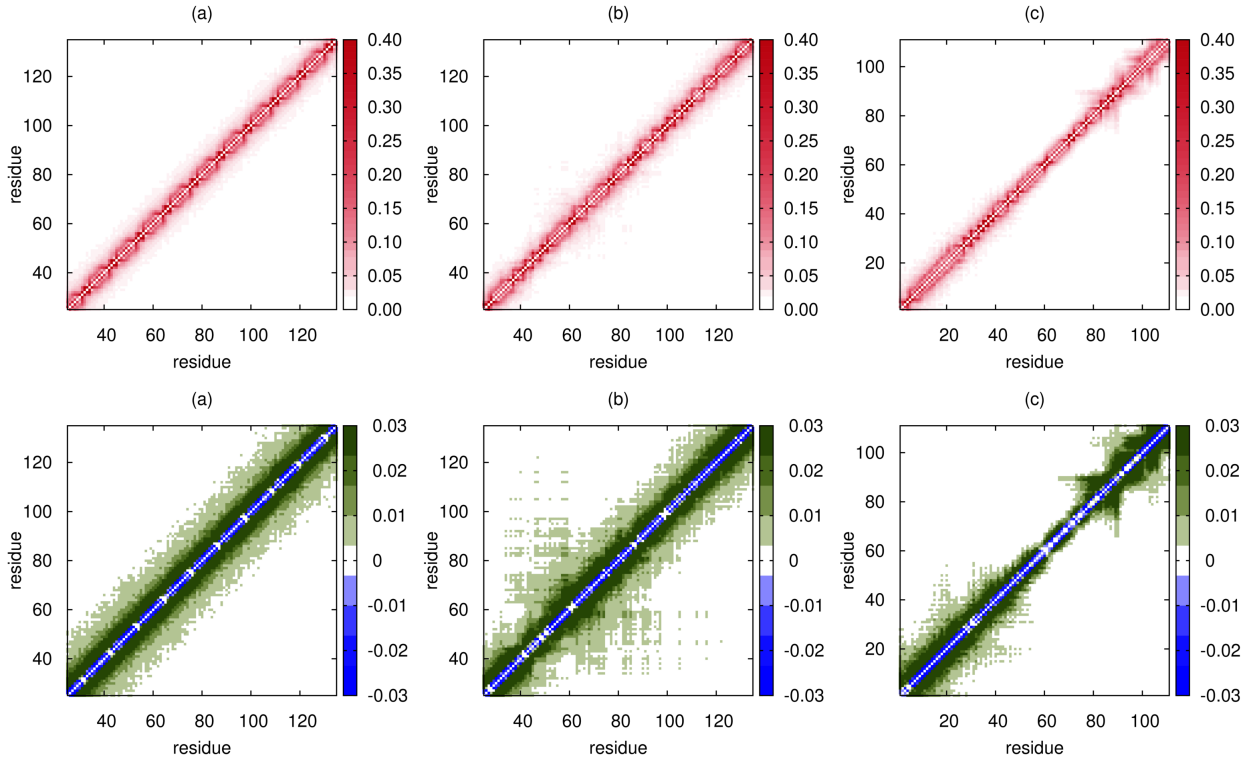


Figure 8: Contact maps for (a) An16, (b) the K25 Tau construct, and (c) Prothymosin-α. The top panel shows the contact maps for the three IDPs calculated using the SOP-IDP simulations. To illustrate the deviations for the corresponding RC contact map we plot the difference map using Eq. (1) (bottom panel). While the contact map for An16 reveals no sequence driven propensities for local structures, the latter two clearly highlight the presence of local domains that are more compact compared to the rest of the sequence. The nearly RC-like behavior of the mid-segment of Prothymosin-α, in stark contrast to the other IDPs studied (see also Figures S9-S11 in the SI), rationalizes the peripheral contribution of compact conformations observed for it in the hierarchical clustering of the conformational ensemble. Each of the contact maps depict a segment of 111 residues (N_T for Prothymosin-α) for optimal visual comparison. The full ranges for (a) and (b) are shown in Figure S11 in the SI.

to an equivalent RC polymer. This can readily be understood as the consequence of the non-specific attractive interactions between non-bonded residues, with sequence specific interactions (final term in Eq. 2) playing a relatively minor role. Without the formation of stable local structures, such enhanced contacts do not alter the RC-like nature of the polymer, as also evidenced by the RC-like $P(R_{ee})$ distribution for An16 (Figure 4).

K25: The results in Figure 8 (b) show that the K25 Tau construct contains a locally compact segment, with preferentially enhanced contacts compared to the rest of the sequence. The segment with locally enhanced contacts is also present in the WT Tau sequence (hTau40), as well as the K25, K23, K44, and hTau23 constructs (shown in Figures S11). We infer that the presence of this region, which results in the formation of the locally preferred structures, leads to the observed deviations from the RC-like behavior for the five sequence constructs.

Prothymosin-α: In this IDP (Figure 8 (c)), the two termini adopt locally compact structures, while the mid-segment of the IDP closely conforms to the RC-limit. For this IDP, electrostatic interactions dominate the conformational properties, and the termini, with both positive- and negatively charged residues, undergo local compaction. However, both the termini are overall negative. Therefore, unfavorable non-local electrostatic interactions prevent the formation of globally compact conformations, resulting in the predominance of extended and semi-extended conformations in the conformational ensemble (Figure 7).

Structural ensemble of α-Synuclein: Our simulations provide considerable insights into the structural details of α-Synuclein, which has been extensively investigated using a variety of experimental methods. The calculated hydrodynamic radius S2 reported in Table S2 ($R_h = 2.89 \pm 0.17$) is in very good agreement with experimental value ($R_h = 2.74 \pm 0.04$) obtained using pulse field gradient NMR.[59] Similarly, the calculated R_g value is in close agreement with experimental estimate (see Table S2), with the difference between simulation and experimental estimates being about 12%. More importantly, our simulations provide a complete characterization of the subtle structural details that are difficult to glean

from experiments. First, the segregation of the conformational landscape for α-Synuclein is in accord with recent experimental works[60–63] that hinted at the presence of distinct conformational states in the structural ensemble, with compact structures having the highest population at equilibrium. Specifically, the difference contact map (Figure 9) reveals that there are long range tertiary interactions between residues in the acidic C-terminal region, and the NAC domain in α-Synuclein. Previous works[64,65] suggest that such interactions, which protect the NAC region against fibrillation, play a key role in impeding aggregation. Interestingly, tertiary contacts between residues in the neighborhood of 74 and 94, first noted in experiments using electron transfer,[66,67] are captured in the contact maps presented in Figure 9. Because of large conformational fluctuations such tertiary interactions, which could serve as initiating nucleation sites for aggregation, the probability of forming such contacts is not substantial but detectable.

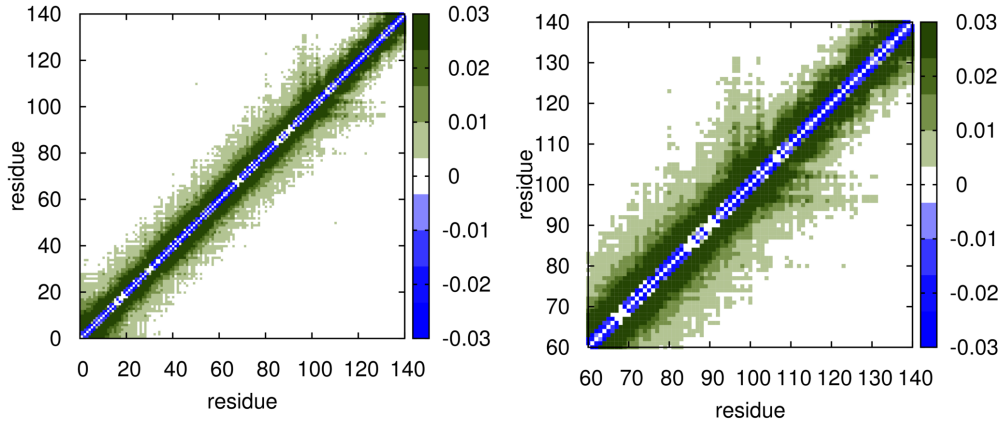


Figure 9: Difference contact maps for the α-Synuclein. Left: whole sequence Right: zoomed-in view highlighting the contacts between the C-terminal and the NAC regions.

Sequence-specific deviations from the RC-limit highlight the heterogeneity in the population of IDPs. We reiterate that it cannot be trivially ascertained that water (or solution conditions used in experiments) act uniformly as a good solvent over the entire peptide sequence of an IDP. To classify IDP sequences as polymers in a good solvent, based solely on estimates of the Flory exponent ν with modest variations in N_T is erroneous, which severely undermines the heterogeneity of their conformations.

Conclusions

Given the multifarious roles that IDPs play in a variety of cellular functions, there is an urgent need to decipher the physical principles governing IDP structure, and dynamics, at the molecular level. We have introduced a robust transferable coarse-grained model for IDPs (SOP-IDP), which quantitatively produces in unprecedented detail agreement with experimental SAXS profiles for IDP sequences of varying composition, and charge densities. Although globally the sizes of the IDPs follow the well known Flory scaling law obtained for synthetic polymers, equilibrium populations of different substates strongly depend on the precise IDP sequence. In the conformational substates of IDPs, we find evidence for specific weak intramolecular interactions between certain residues, giving rise to population of conformations with local structures. Our studies also reveal that the adaptability of IDPs arising from their conformational plasticity, which cannot be anticipated using sequence characteristics (such as fraction of positive and negatively charged residues) alone,[27] is important in their ability to interact with a multitude of partners to form fuzzy complexes stabilized by specific intermolecular interactions. In other words, we propose that the observed sequence-dependent conformational heterogeneity determines the functions of IDPs. The accuracy of the computations in describing the equilibrium ensemble of a number of IDPs sets the stage for describing IDP-RNA interactions, which is needed to understand ribonucleoprotein stress granule formation in eukaryotes, as well as quantitative descriptions of the formation of biomolecular condensates at the molecular level.

Methods

Development of the Model: In the Self-Organized Polymer (SOP-IDP) model, except for glycine and alanine, the amino acid residues are represented using a backbone and a side chain (SC) bead. We use the C_α atoms to model glycine and alanine owing to their small side chains. The radii of the individual beads are given in Table S1 of the SI. The energy

function in the SOP-IDP model is:

$$
\begin{aligned}
E &= -\sum_{i=1}^{N_B} \frac{k}{2} R_0^2 \log\left(1 - \frac{(r_i - r_{ref,i})^2}{R_0^2}\right) + \sum_{i=1}^{N_{loc}} \epsilon_{loc} \left(\frac{\sigma_i}{r_i}\right)^6 + \sum_{i,j}{}' \frac{e_i e_j \exp\left(-\kappa r_{ij}\right)}{\varepsilon r_{ij}} \\
&+ \sum_{i=1}^{N_{BB}} \epsilon_{BB} \left[\left(\frac{\sigma_i}{r_i}\right)^{12} - 2\left(\frac{\sigma_i}{r_i}\right)^6\right] + \sum_{i=1}^{N_{BS}} \epsilon_{BS} \left[\left(\frac{\sigma_i}{r_i}\right)^{12} - 2\left(\frac{\sigma_i}{r_i}\right)^6\right] \\
&+ \sum_{i=1}^{N_{SS}} \epsilon_{SS} \left|\epsilon_i - 0.7\right| \left[\left(\frac{\sigma_i}{r_i}\right)^{12} - 2\left(\frac{\sigma_i}{r_i}\right)^6\right] \qquad (2)
\end{aligned}
$$

The first term in Eq. 2, for a chain with N_B bonds, represents bonded interactions, which are described using a finite non-linear elastic (FENE) potential. The purely repulsive second term acts only between bead pairs (total number N_{loc}) that are not covalently bonded, but belong to residues separated by ≤ 2 along the polypeptide chain. It represents the excluded volume interactions that prevent unphysical bond-crossing between pairs of beads that closely follow one another along the polymer chain. The third term is a screened Coulomb potential, that accounts for electrostatic interactions between all pairs of charged residues. The charge corresponding to a residue is assigned to the SC. The titratable residue histidine is treated as neutral in the model, unless it is experimentally determined to be in the protonated state. The parameters κ and ε in the third term in Eq. 2 are the inverse Debye length, and the dielectric constant, respectively. These three terms were previously used in protein folding and related studies,[68–70] where the justifications for their choices are given.

The final three terms in Eq. 2 represent interactions between backbone–backbone (BB), backbone–side chain (BS), and side chain–side chain (SS) beads, respectively. The total numbers of such BB, BS, and SS pairs, are respectively, N_{BB}, N_{BS} and N_{SS}. Interactions between the side chains depend only on the identities of the amino acids without bias towards any specific structure. The parameter ϵ_i in the final term in Eq. 2 is obtained from the Betancourt-Thirumalai statistical potential.[71] The ϵ_i values are unique to each pair of amino acid residues, and hence, depend explicitly on the IDP sequence. In Eq. 2, the three

parameters ϵ_{BB}, ϵ_{BS} and ϵ_{SS} set the energy scales corresponding to the non-local interactions, and are the only free parameters in the energy function. Following the standard convention used in polymer physics, these three parameters are expressed in units of k_BT (with T=298 K). We determine the initial values of ϵ_{BB}, ϵ_{BS} and ϵ_{SS} in a top-down fashion, by using SAXS data (R_g, and $I(q)$)for only three short IDPs ($24 \leq N_T \leq 131$) as constraints in the parametrization scheme. Subsequently, to find the optimum values of the parameters, only the experimental R_g estimates for three long IDP sequences ($202 \leq N_T \leq 441$) were used as constraints in the parametrization scheme. Further details of this learning procedure are provided in the supporting information.

Simulations: We used the LAMMPS molecular dynamics simulator for all the simulations,[72] which were performed using the under-damped Langevin dynamics,.[73] The equations of motion were integrated with a time step of 30 fs. To obtain statistically meaningful results, we carried out ten independent simulations for each system, which we ascertained are sufficient to obtain converged results. Each trajectory was equilibrated for 10^8 simulation steps, following which the production runs were carried out for 2×10^8 simulation steps for sequence lengths < 130, and 5×10^8 simulation steps for longer sequences. Visual molecular dynamics (VMD) was used for visualization, as well as some data analyses.[74]

In our simulations, all the non-bonded interaction terms in Eq. 2 were implemented in their truncated and shifted forms, with the cut-off separation for short-range interactions set at 2.4 nm. The cut-off separation for the screened Coulomb potential was consistently chosen to be greater than $4\kappa^{-1}$. The inverse Debye length (κ) was determined from the corresponding monovalent salt concentrations employed in the simulations. We set the dielectric constant, $\varepsilon = 78$.

Data Analyses: The simulated SAXS profiles were computed using the Debye formula,

$$I(q) = \sum_{i=1}^{N_{tot}} \sum_{j=1}^{N_{tot}} f_i(q) f_j(q) \frac{\sin(qr_{ij})}{qr_{ij}}, \tag{3}$$

with q-dependent structure factors ($f(q)$), which were reported elsewhere.[75] In Eq. 3, N_{tot} is the total number of beads in a given IDP.

Hydrodynamic radius, R_h: Dynamic light scattering (DLS), and fluorescence correlation spectroscopy (FCS) experiments[9,76–78] are routinely used to measure R_h. The value for R_h for a polymer is calculated as the radius of a hard sphere that has the same effective diffusive behavior as the polymer. We computed R_h from the simulations using:

$$\frac{1}{R_h} = \frac{1}{N_{tot}^2} \left\langle \sum_{i \neq j} \frac{1}{r_{ij}} \right\rangle, \tag{4}$$

where N_{tot} is the total number of beads in the IDP model, and r_{ij} is the distance between beads j and i. The angular braces denote the ensemble average.

Shape Parameters: We use Δ and S to characterize the shapes of folded and unfolded proteins,[79,80] which are defined as,

$$\Delta = \frac{3}{2} \frac{\sum_{i=1}^{3} \left(\lambda_i - \bar{\lambda}\right)^2}{\left(3\bar{\lambda}\right)^2}, \; S = \frac{\prod_{i=1}^{3} \left(\lambda_i - \bar{\lambda}\right)}{\bar{\lambda}^3}, \tag{5}$$

where λ_i are the eigenvalues of the gyration tensor, and $\bar{\lambda} = (\lambda_1 + \lambda_2 + \lambda_3)/3$.[79] By definition, $R_g^2 = 3\bar{\lambda}$. The Δ parameter characterizes the asphericity of conformations, and is bound between 0 (perfectly spherical) and 1 (linear). The shape parameter S, which is negative (positive) for oblate (prolate) ellipsoids satisfies the bound $-\frac{1}{4} \leq S \leq 2$.[81]

Hierarchical Clustering: To identify representative conformations populating the IDP ensembles, as well as characterize their structural heterogeneity, we performed hierarchical clustering[82] using a pairwise distance metric D_{ij}, defined as,

$$D_{ij} = \left(\frac{1}{N_{pairs}} \sum_{a,b} |(r_{a,b}^i - r_{a,b}^j)|^2 \right)^{\frac{1}{2}} \tag{6}$$

where $r_{a,b}^i$ and $r_{a,b}^j$ are the pairwise distances between the C_α atoms a and b, in snapshots i and j, respectively; N_{pairs} is the total number of C_α pairs. The terminal C_α atoms were

excluded in the evaluation of the distance matrix. The Ward variance minimization criterion,[83] as available within the *scipy* module, was employed to identify the distinct clusters. The hierarchical organization of clusters was visualized in the form of dendrograms. The structure exhibiting the lowest root-mean-square-deviation (RMSD) with respect to all the other members is identified as the representative structure of a given cluster.

Acknowledgement

The research was supported by the National Science Foundation (CHE 16-36424), the Collie Welch Regents Chair (F-0019), and the National Institutes of Health (R01 GM107703). We are grateful to the Texas Advanced Computing Center (TACC) for providing the needed computing time.

Supporting Information Available

Details of the learning procedure used to calibrate the SOP-IDP model; Distributions of the shape parameters for all the 24 IDP sequences; End-to-end (R_{ee}) distributions for all the 24 IDP sequences; Sequence compositional properties of the IDP sequences; Contact maps for selected IDP sequences; FASTA representation of the IDP sequences.

This material is available free of charge via the Internet at `http://pubs.acs.org/`.

References

(1) Babu, M. M.; van der Lee, R.; de Groot, N. S.; Gsponer, J. *Curr. Opin. Struct. Biol.* **2011**, *21*, 432–440.

(2) Wright, P. E.; Dyson, H. J. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18.

(3) Uversky, V. N.; Oldfield, C. J.; Dunker, A. K. *Annu. Rev. Biophys.* **2008**, *37*, 215–246.

(4) Babu, M. M. *Biochem. Soc. Trans.* **2016**, *44*, 1185–1200.

(5) Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradović, Z.; Dunker, A. K. *J. Mol. Biol.* **2002**, *323*, 573 – 584.

(6) Gsponer, J.; Babu, M. M. *Cell Rep.* **2012**, *2*, 1425–1437.

(7) Tompa, P. *Trends Biochem. Sci.* **2012**, *37*, 509 – 516.

(8) Hyman, A. A.; Weber, C. A.; Jülicher, F. *Annu. Rev. Cell Dev. Biol.* **2014**, *30*, 39–58.

(9) Busch, D. J.; Houser, J. R.; Hayden, C. C.; Sherman, M. B.; Lafer, E. M.; Stachowiak, J. C. *Nat. Commun.* **2015**, *6*, 7875.

(10) Dima, R. I.; Thirumalai, D. *Bioinformatics* **2004**, *20*, 2345–2354.

(11) Shin, Y.; Brangwynne, C. P. *Science* **2017**, *357*, eaaf4382.

(12) Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31–38.

(13) de Sancho, D.; Baker, C. M.; Best, R. B. *Computational Approaches to Protein Dynamics: From Quantum to Coarse-Grained Methods* **2014**, 283.

(14) Brangwynne, C. P.; Tompa, P.; Pappu, R. V. *Nat. Phys.* **2015**, *11*, 899–904.

(15) Dignon, G. L.; Zheng, W.; Kim, Y. C.; Best, R. B.; Mittal, J. *PLoS Comput. Biol.* **2018**, *14*, e1005941.

(16) Liu, Y.; Wang, X.; Liu, B. *Brief. Bioinform.* **2017**, bbx126.

(17) Levine, Z. A.; Shea, J.-E. *Curr. Opin. Struct. Biol.* **2017**, *43*, 95 – 103.

(18) Potenza, E.; Domenico, T. D.; Walsh, I.; Tosatto, S. C. *Nucleic Acids Res.* **2015**, *43*, D315–D320.

(19) Das, R. K.; Ruff, K. M.; Pappu, R. V. *Curr. Opin. Struct. Biol.* **2015**, *32*, 102 – 112.

(20) Wright, P. E.; Dyson, H. J. *J. Mol. Biol.* **1999**, *293*, 321 – 331.

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

(22) Kohn, J. E.; Millet, I. S.; Jacob, J.; Zagrovic, B.; Dillon, T. M.; Cingel, N.; Dothagar, R. S.; Seifert, S.; Thiyagarajan, P.; Sosnick, T. R.; Hasan, M. Z.; Pande, V. S.; Ruczinski, I.; Doniach, S.; Plaxco, K. W. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 12491–12496.

(23) Wilson, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. *Biochemistry* **1999**, *38*, 16424–16431.

(24) Receveur-Bréchot, V.; Durand, D. *Curr. Protein Pep. Sci.* **2012**, *13*, 55–75.

(25) Tamiola, K.; Acar, B.; Mulder, F. A. *J. Am. Chem. Soc.* **2010**, *132*, 18000–18003.

(26) Samanta, H. S.; Zhuravlev, P. I.; Hinczewski, M.; Hori, N.; Chakrabarti, S.; Thirumalai, D. *Soft Matter* **2017**, *13*, 3622–3638.

(27) Samanta, H. S.; Chakraborty, D.; Thirumalai, D. *J. Chem. Phys.* **2018**, *149*, 163323.

(28) Bernado, P.; Svergun, D. I. *Mol. BioSyst.* **2012**, *8*, 151–167.

(29) Gibbs, E. B.; Cook, E. C.; Showalter, S. A. *Arch. Biochem. Biophys.* **2017**, *628*, 57 – 70.

(30) Brucale, M.; Schuler, B.; Samorì, B. *Chem. Rev.* **2014**, *114*, 3281–3317.

(31) Fu, B.; Vendruscolo, M. *Intrinsically Disordered Proteins Studied by NMR Spectroscopy*; Springer International Publishing: Cham, 2015; pp 35–48.

(32) Doi, M.; Edwards, S. F. *The theory of polymer dynamics*; oxford university press, 1988; Vol. 73.

(33) Kikhney, A. G.; Svergun, D. I. *FEBS Lett.* **2015**, *589*, 2570–2577.

(34) Piana, S.; Klepeis, J. L.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98–105.

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

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

(37) Henriques, J.; Cragnell, C.; Skepö, M. *J. Chem. Theory Comput.* **2015**, *11*, 3420–3431.

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

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

(40) Cragnell, C.; Rieloff, E.; Skepö, M. *J. Mol. Biol.* **2018**, *430*, 2478–2492.

(41) Wu, H.; Wolynes, P. G.; Papoian, G. *J. Phys. Chem. B*

(42) Cragnell, C.; Durand, D.; Cabane, B.; Skepö, M. *Proteins: Struct., Funct., Bioinf.* **2016**, *84*, 777–791.

(43) Kjaergaard, M.; Nørholm, A.-B.; Hendus-Altenburger, R.; Pedersen, S. F.; Poulsen, F. M.; Kragelund, B. B. *Prot. Sci.* **2010**, *19*, 1555–1564.

(44) Mercadante, D.; Milles, S.; Fuertes, G.; Svergun, D. I.; Lemke, E. A.; Grter, F. *J. Phys. Chem. B* **2015**, *119*, 7975–7984.

(45) Arbes, M.; Maffei, M.; Cordeiro, T. N.; Teixeira, J. M.; Prez, Y.; Bernad, P.; Roche, S.; Pons, M. *Structure* **2017**, *25*, 630 – 640.e4.

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

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

(48) Uversky, V. N.; Gillespie, J. R.; Millett, I. S.; Khodyakova, A. V.; Vasilenko, R. N.; Vasiliev, A. M.; Rodionov, I. L.; Kozlovskaya, G. D.; Dolgikh, D. A.; Fink, A. L.; Doniach, S.; Permyakov, E. A.; Abramov, V. M. *Biochem. Biophys. Res. Commun.* **2000**, *267*, 663 – 668.

(49) Lens, Z.; Dewitte, F.; Mont, D.; Baert, J.-L.; Bompard, C.; Snchal, M.; Lint, C. V.; de Launoit, Y.; Villeret, V.; Verger, A. *Biochem. Biophys. Res. Commun.* **2010**, *399*, 104 – 110.

(50) Schwalbe, M.; Ozenne, V.; Bibow, S.; Jaremko, M.; Jaremko, L.; Gajda, M.; Jensen, M. R.; Biernat, J.; Becker, S.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *Structure* **2014**, *22*, 238 – 249.

(51) Nairn, K. M.; Lyons, R. E.; Mulder, R. J.; Mudie, S. T.; Cookson, D. J.; Lesieur, E.; Kim, M.; Lau, D.; Scholes, F. H.; Elvin, C. M. *Biophys. J.* **2008**, *95*, 3358 – 3365.

(52) Holt, C.; Srensen, E. S.; Clegg, R. A. *FEBS J.* **2009**, *276*, 2308–2323.

(53) Henriques, J.; Skepö, M. *J. Chem. Theory Comput.* **2016**, *12*, 3407–3415.

(54) Mylonas, E.; Hascher, A.; Bernadó, P.; Blackledge, M.; Mandelkow, E.; Svergun, D. I. *Biochemistry* **2008**, *47*, 10345–10353.

(55) Teraoka, I. Polymer Solutions: an Introduction to Physical Properties 2002.

(56) Rubinstein, M.; Colby, R. H. *Polymer physics*; Oxford University Press, 2015.

(57) Honeycutt, J. D.; Thirumalai, D. *J. Chem. Phys.* **1989**, *90*, 4542–4559.

(58) Zambelli, A. E. *F100Res.* **2016**, *5*, 2809.

(59) Sevcsik, E.; Trexler, A. J.; Dunn, J. M.; Rhoades, E. *J. Am. Chem. Soc.* **2011**, *133*, 7152–7158.

(60) Frimpong, A. K.; Abzalimov, R. R.; Uversky, V. N.; Kaltashov, I. A. *Proteins* **2010**, *78*, 714–722.

(61) Nath, A.; Sammalkorpi, M.; DeWitt, D. C.; Trexler, A. J.; Elbaum-Garfinkle, S.; O'Hern, C. S.; Rhoades, E. *Biophys. J.* **2012**, *103*, 1940–1949.

(62) UIlman, O.; Fisher, C. K.; Stultz, C. M. *J. Am. Chem. Soc.* **2011**, *133*, 19536–19546.

(63) Dedmon, M. M.; Lindorff-Larsen, K.; Christodoulou, J.; Vendruscolo, M.; Dobson, C. M. *J. Am. Chem. Soc.* **2005**, *127*, 476–477.

(64) Hoyer, W.; Cherny, D.; Subramaniam, V.; Jovin, T. M. *Biochemistry* **2004**, *43*, 16233–16242.

(65) Trexler, A. J.; Rhoades, E. *Biophys J.* **2010**, *99*, 3048–3055.

(66) Lee, J. C.; Gray, H. B.; Winkler, J. R. *J. Am. Chem. Soc.* **2005**, *127*, 16388–16389.

(67) Lee, J. C.; Lai, B. T.; Kozak, J. J.; Gray, H. B.; Winkler, J. R. *J. Phys. Chem. B.* **2007**, *111*, 2107–2112.

(68) Reddy, G.; Thirumalai, D. *J. Phys. Chem. B* **2015**, *119*, 11358–11370.

(69) Liu, Z.; Reddy, G.; Thirumalai, D. *J. Phys. Chem. B* **2012**, *116*, 6707–6716.

(70) Tehver, R.; Thirumalai, D. *Structure* **2010**, *18*, 471 – 481.

(71) Betancourt, M. R.; Thirumalai, D. *Prot. Sci.* **1999**, *8*, 361–369.

(72) Plimpton, S. *J. Comp. Phys.* **1995**, *117*, 1 – 19.

(73) Schneider, T.; Stoll, E. *Phys. Rev. B* **1978**, *17*, 1302–1322.

(74) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph Model.* **1996**, *14*, 33 – 38.

(75) Tong, D.; Yang, S.; Lu, L. *J. Appl. Crystallogr.* **2016**, *49*, 1148–1161.

(76) Stetefeld, J.; McKenna, S. A.; Patel, T. R. *Biophys. Rev.* **2016**, *8*, 409–427.

(77) Sherman, E.; Itkin, A.; Kuttner, Y. Y.; Rhoades, E.; Amir, D.; Haas, E.; Haran, G. *Biophys. J.* **2008**, *94*, 4819 – 4827.

(78) Gast, K.; Fiedler, C. *Intrinsically disordered protein analysis*; Springer, 2012; pp 137–161.

(79) Hyeon, C.; Dima, R. I.; Thirumalai, D. *J. Chem. Phys.* **2006**, *125*, 194905.

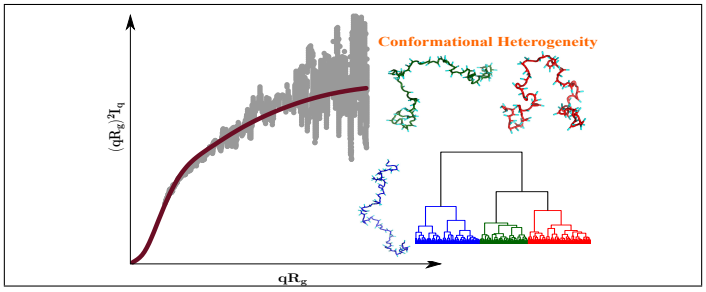
(80) Dima, R. I.; Thirumalai, D. *J. Phys. Chem. B* **2004**, *108*, 6564–6570.

(81) Aronovitz, J. A.; Nelson, D. R. *J. Physique* **1986**, *47*, 1445–1456.

(82) Maimon, O., Rokach, L., Eds. *Data Mining and Knowledge Discovery Handbook*; Springer, 2010.

(83) Ward, J. H. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244.

Graphical TOC Entry



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 transferable model for accurately describing not only the average properties of IDPs but also the details of the structural ensemble. To this end, we created the SOP-IDP model, which is built on the successful Self-Organized Polymer (SOP) model used to investigate 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, 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 Waal 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 (in the Kratky representation) 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.[3,4] Histatin-5 is a small (24 residue) IDP that has often been used as reference for testing the validity of computational models.[3,5] The other two IDPs, ACTR ($N_T = 71$) and hNHE1 ($N_T = 131$) are also reasonably well studied.[4,6] The choice of these three IDP sequences enabled us to explore a broad range in N_T in order to obtain our initial estimate of the unknown parameters in our model. The SAXS profiles for Histatin-5, ACTR and hNHE1 at 150 mM monovalent salt concentration have low noise-to-signal ratio, which makes it objective comparisons feasible with the simulated profiles at equivalent ionic strengths.

By considering the simultaneous reproduction of the experimental R_g values, and scattering profiles in the low q regime, we found that the parameter set, $\epsilon_{BB} = 0.2$, $\epsilon_{BS} = 0.1$, and $\epsilon_{SS} = 0.3$ is optimal for describing the three sequences in the training set. The Kratky plots obtained from the simulations overlaid on the experimental profiles in Figure S1 (a)-(c), show excellent agreement. However, we observed that for long IDP sequences, the initial parameter set results in systematic overestimation of R_g values (Figure S1). This deviation indicates that the interactions determining the compaction of IDPs, are not strong enough to counterbalance the increased conformational entropy for long IDP sequences. Further optimization of the parameter set was thus necessary, and was carried out using R_g values from experiments for three long IDP sequences, namely K32, K23, and hTau40 as constraints. Taking a cue from our learning procedure, which suggested that in the neighborhood of the learnt parameters the simulated R_g values decrease monotonically with the interaction strengths (Figure S2), most notably ϵ_{BS}, we systematically varied the backbone-side chain

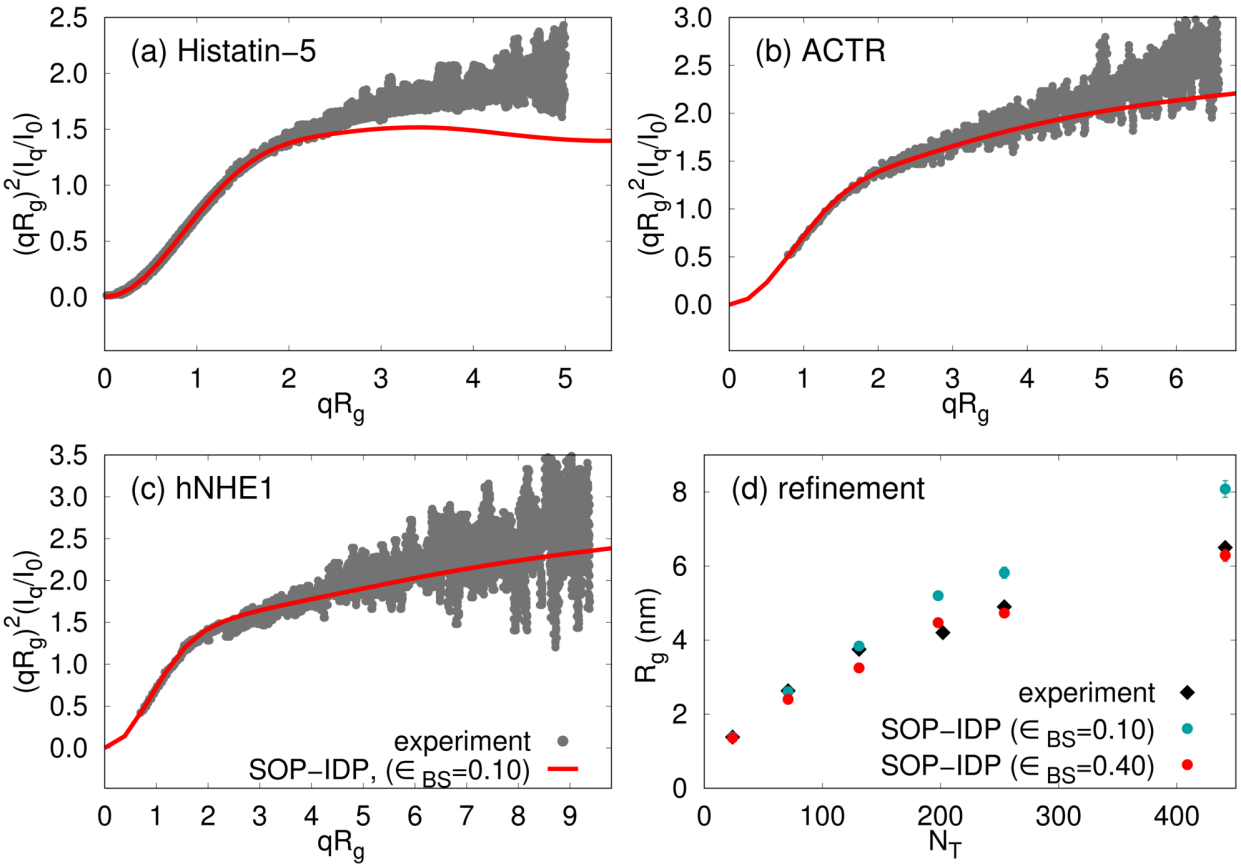


Figure S1: SAXS profiles in the Kratky representation (red curves), computed from SOP-IDP model simulations using the learned parameters (ϵ_{BB}, ϵ_{BS}, ϵ_{SS}) = (0.2, 0.1, 0.3); for (a) Histatin-5, (b) ACTR, and (c) hNHE1. The energy scales are in units of k_BT, with $T = 298K$, Experimental SAXS profiles are shown with gray dots, and are from references.[3,4] The value of ϵ_{BS} is further refined to obtain better agreement with experimental R_g values for long sequences. Note that even without the additional refinement ($\epsilon_{BS} = 0.1$), the deviation form experiments is relatively small.

interaction strength, keeping the other two parameters fixed. We found that rescaling ϵ_{BS} to 0.4 was sufficient to quantitatively reproduce the experimental R_g values over the entire range in chain length $24 \leq N_T \leq 441$ (Figure S1 (d)). The SAXS profiles corresponding to the original training set could still be quantitatively reproduced after rescaling ϵ_{BS} (shown in the manuscript, Figure 1). All the results described in this work are based on the refined parameter set $\epsilon_{BB} = 0.2$, $\epsilon_{BS} = 0.4$, and $\epsilon_{SS} = 0.3$.

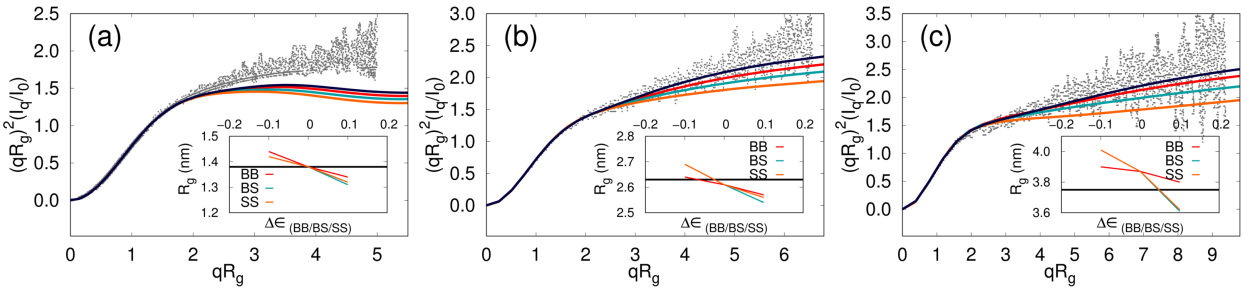


Figure S2: Schematics showing some results used to arrive at initial estimates for the free parameters in the SOP-IDP model, using (a) Histatin-5, (b) ACTR, and (c) hNHE1. Black, red, cyan, and orange curves in the main plots show SAXS profiles in the Kratky representation, obtained using parameter sets (0.1, 0.1, 0.2), (0.2, 0.1, 0.3), (0.2, 0.1, 0.4) and (0.3, 0.2, 0.4) respectively for (ϵ_{BB}, ϵ_{BS}, ϵ_{SS}). The inset plots in all figures show the changes in the simulated R_g values, when a single parameter ϵ_{BB}/ ϵ_{BS}/ ϵ_{SS} is varied while keeping the other two fixed at their learnt values. $\Delta\epsilon$ represents the deviation from the corresponding learnt parameter value. For example, the red line is obtained by fixing ϵ_{BS} and ϵ_{SS} while varying ϵ_{BB}. The horizontal lines in the inset plots represent experimental R_g values. In the immediate vicinity of the estimated parameters (0.2, 0.1, 0.3), the changes with respect to all parameters are monotonic, and R_g is found to decrease with increased interaction strengths. Experimental results are taken from previous studies.[3,4]

Distributions of shape parameters

In Figures S3 - S6, we show the distributions of shape parameters Δ and S (see Eq. 5 in main text) for all 24 IDP sequences. Remarkably, the results in Figures S3 - S6 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 S2), which merely shows that the average Δ and S do not elucidate the granularity of the IDP conformations.

Table S2: Conformational properties of IDPs. The values of R_g, R_{ee} and R_h are in nm. The numbers in parenthesis represent standard errors. The reasons for large errors in the mean Δ and S are explained in the SI text.

IDP	N_T	R_g (exp)	R_g (sim)	R_{ee} (sim)	R_h (sim)	Δ	S
Histatin-5	24	1.38	1.36 (0.09)	3.09 (0.29)	1.32 (0.08)	0.40 (0.16)	0.50 (0.34)
ACTR	71	2.63	2.40 (0.17)	5.62 (0.65)	2.07 (0.10)	0.41 (0.19)	0.49 (0.39)
Nucleoporin	81	2.7	2.59 (0.14)	6.18 (0.71)	2.19 (0.12)	0.40 (0.19)	0.49 (0.41)
SH4-UD	85	2.9	2.78 (0.15)	6.55 (0.75)	2.30 (0.11)	0.43 (0.19)	0.53 (0.43)
Sic1	90	3.21	2.84 (0.16)	6.77 (0.75)	2.36 (0.12)	0.42 (0.19)	0.52 (0.40)
p53	93	2.87	2.95 (0.16)	7.08 (0.80)	2.43 (0.11)	0.42 (0.19)	0.54 (0.40)
Prothymosin α	111	3.79	3.91 (0.21)	9.37 (0.97)	2.92 (0.14)	0.51 (0.19)	0.72 (0.45)
ERM TADn	122	3.81	3.14 (0.22)	7.37 (0.89)	2.63 (0.17)	0.38 (0.19)	0.45 (0.39)
hNHE1	131	3.75	3.25 (0.23)	7.45 (0.93)	2.69 (0.12)	0.39 (0.19)	0.48 (0.39)
α-Synuclein	140	4.01	3.54 (0.23)	8.16 (1.06)	2.89 (0.17)	0.40 (0.19)	0.47 (0.40)
An16	185	5.0	4.32 (0.28)	10.24 (1.21)	3.38 (0.18)	0.43 (0.19)	0.54 (0.41)
Osteopontin	273	5.5	4.72 (0.35)	10.08 (1.39)	3.82 (0.35)	0.40 (0.19)	0.52 (0.40)
K19	99	3.5	2.96 (0.17)	7.00 (0.81)	2.44 (0.09)	0.43 (0.19)	0.54 (0.41)
K18	130	3.8	3.48 (0.21)	8.24 (0.96)	2.80 (0.11)	0.43 (0.19)	0.54 (0.42)
K17	143	3.6	3.74 (0.22)	8.89 (1.03)	2.98 (0.12)	0.43 (0.19)	0.54 (0.42)
K10	167	4.0	3.90 (0.25)	9.08 (1.11)	3.13 (0.12)	0.41 (0.19)	0.50 (0.41)
K27	171	3.7	4.06 (0.25)	9.64 (1.14)	3.20 (0.13)	0.43 (0.19)	0.55 (0.42)
K16	174	3.9	4.18 (0.26)	9.89 (1.16)	3.28 (0.12)	0.42 (0.19)	0.51 (0.40)
K25	185	4.1	3.97 (0.28)	9.03 (1.23)	3.22 (0.12)	0.36 (0.18)	0.41 (0.37)
K32	202	4.15	4.47 (0.28)	10.57 (1.24)	3.49 (0.13)	0.42 (0.19)	0.52 (0.42)
K23	254	4.9	4.73 (0.34)	10.75 (1.41)	3.75 (0.14)	0.38 (0.19)	0.44 (0.39)
K44	283	5.2	5.10 (0.36)	11.63 (1.54)	4.01 (0.16)	0.40 (0.19)	0.49 (0.40)
hTau23	352	5.3	5.71 (0.40)	12.79 (1.68)	4.44 (0.15)	0.42 (0.19)	0.53 (0.40)
hTau40	441	6.5	6.29 (0.48)	14.11 (1.98)	4.88 (0.17)	0.42 (0.19)	0.54 (0.41)

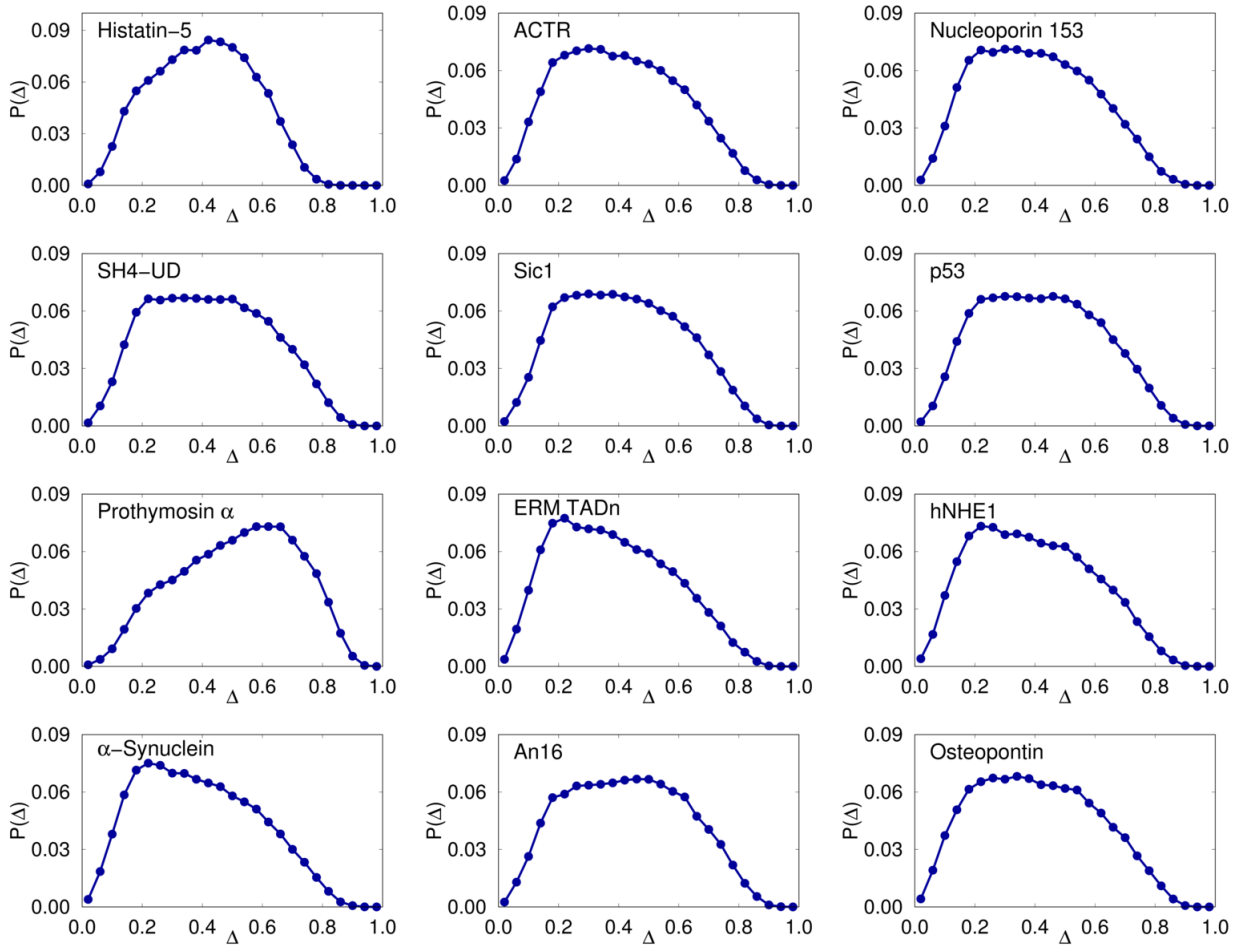


Figure S3: Distributions of the shape parameter Δ (Eq. 5 in the main text), for all non-Tau IDP sequences listed in Table S2.

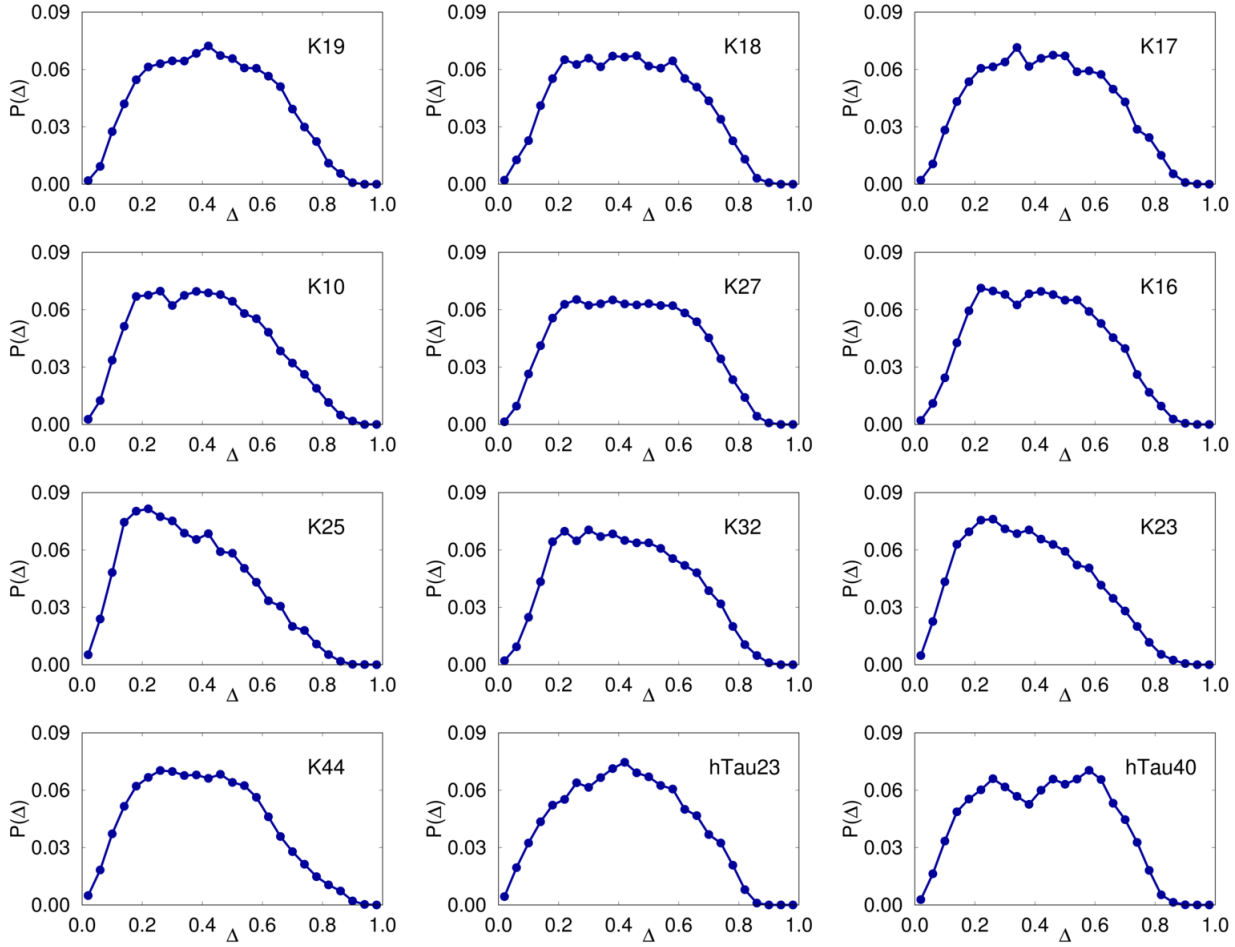


Figure S4: Same as Figure S3, except the results are for the Tau protein constructs.

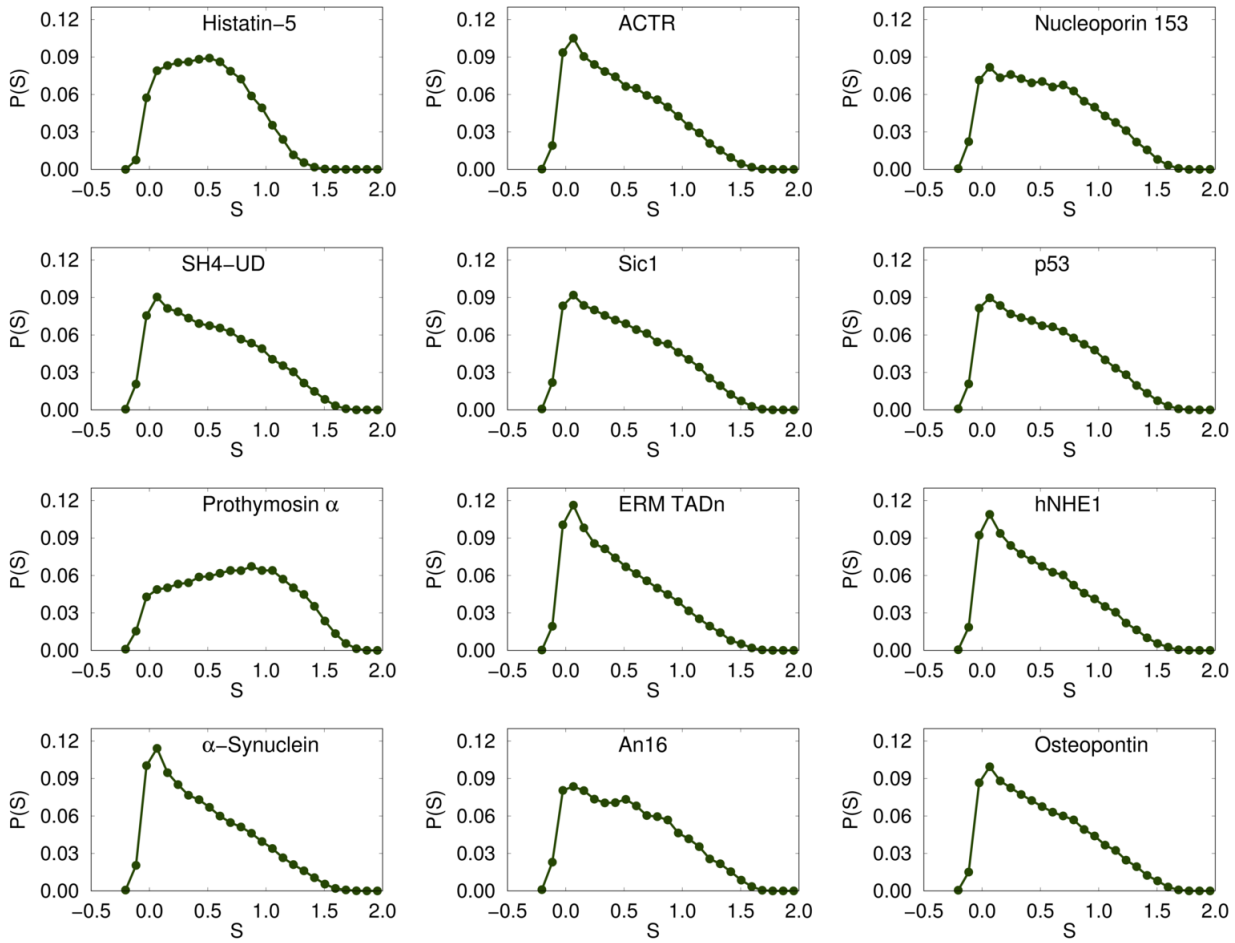


Figure S5: Distributions of the shape parameter S (Eq. 5 in the main text), for the twelve non-Tau IDP sequences.

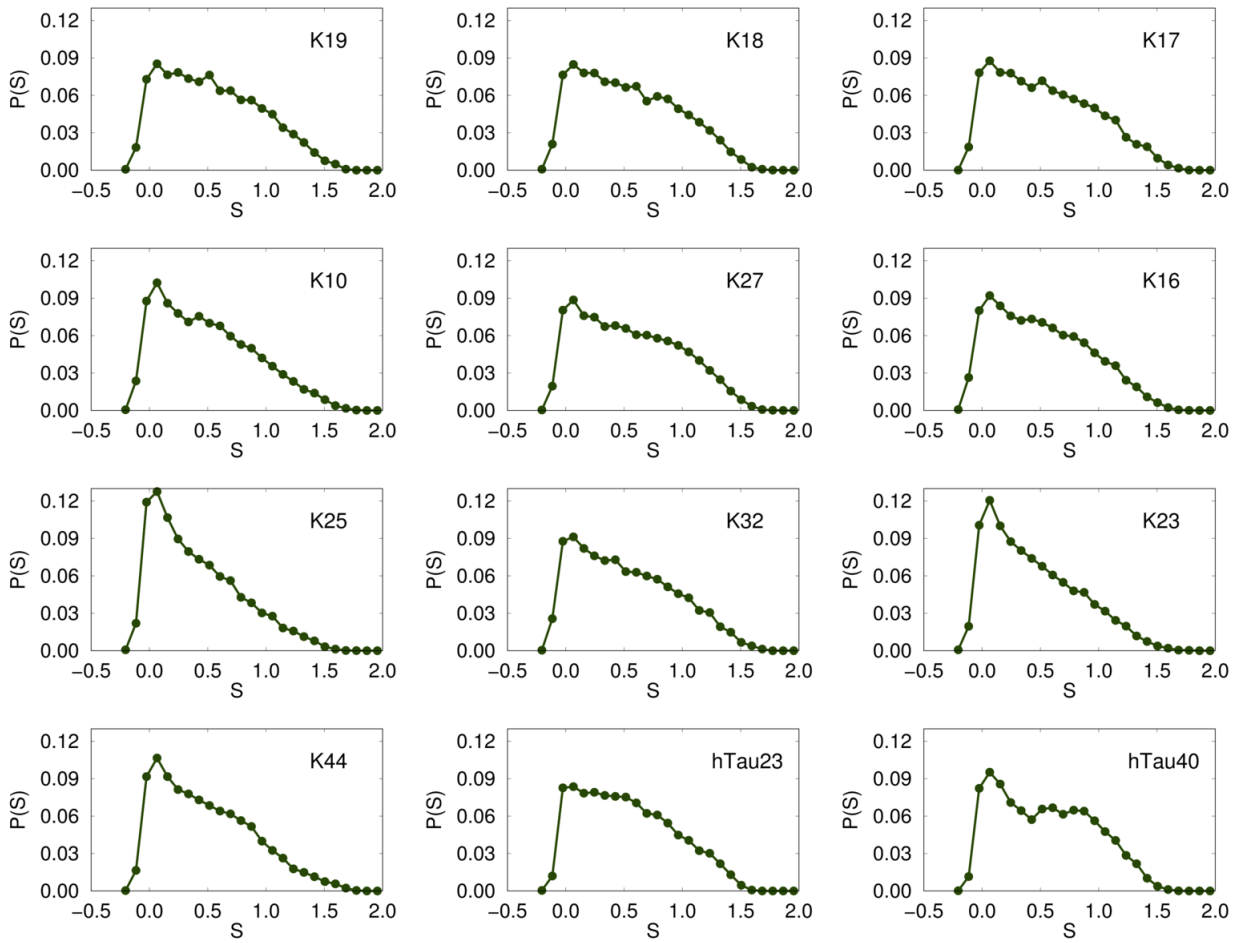


Figure S6: Same as Figure S5, except the results are for the Tau protein constructs.

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

In Figures S7 and S8, we show the R_{ee} distributions for all 24 IDP sequences, and compare them with rigorous theoretical curves for the random coil-like behavior, and the Gaussian chain. The theoretical results for the random coil-like and Gaussian chain is given by[7] $P(x) = Cx^g e^{-\alpha x^\delta}$ 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.[7] 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 S8) 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 S8). This is an example where the apparent solvent quality changes upon increasing N_T.

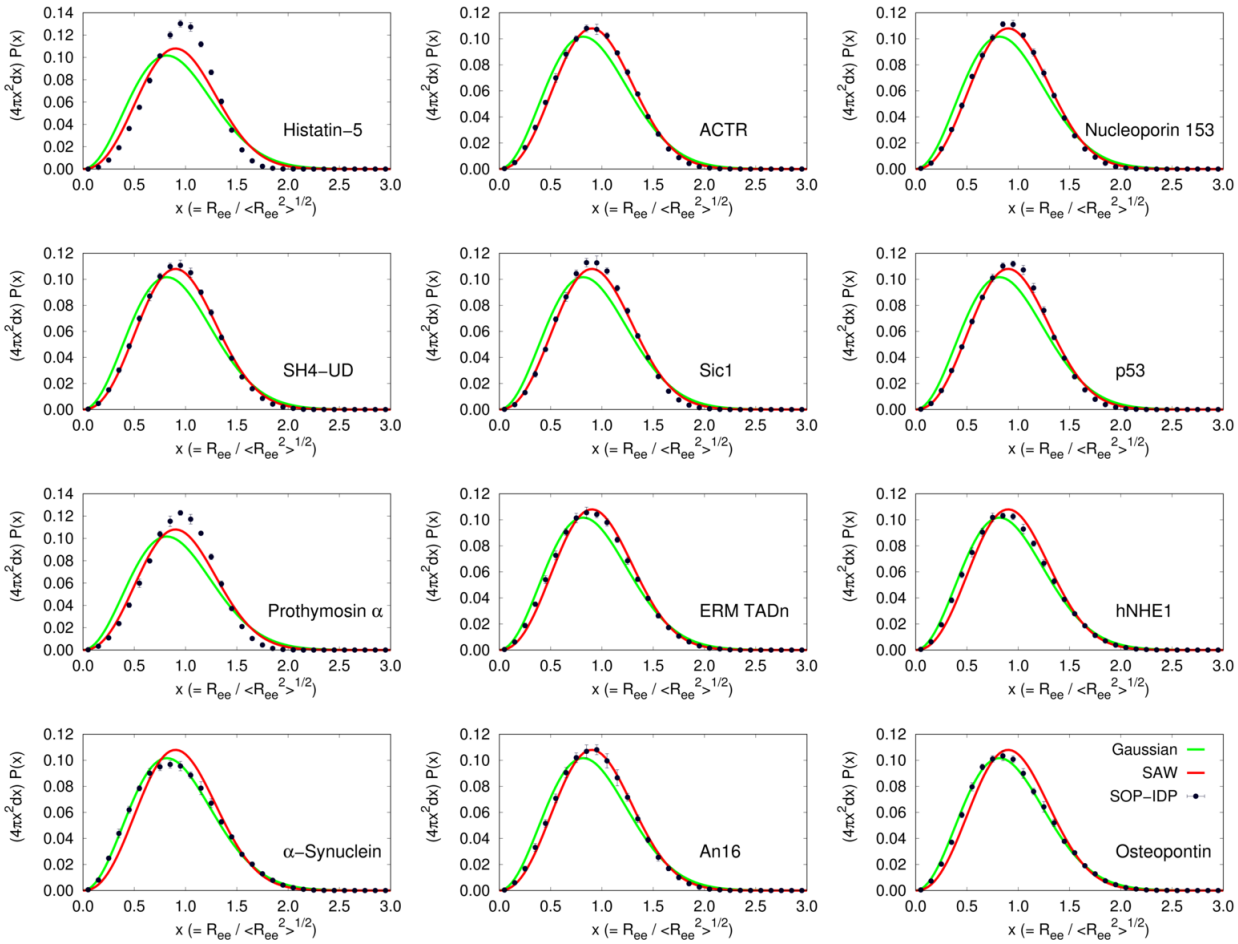


Figure S7: (black points) Plots for R_{ee} (scaled by $\langle R_{ee}^2 \rangle^{\frac{1}{2}}$ distance distributions for non-Tau IDP sequences. Theoretical distributions for random coil-like behavior and Gaussian chains are shown in red and green, respectively.

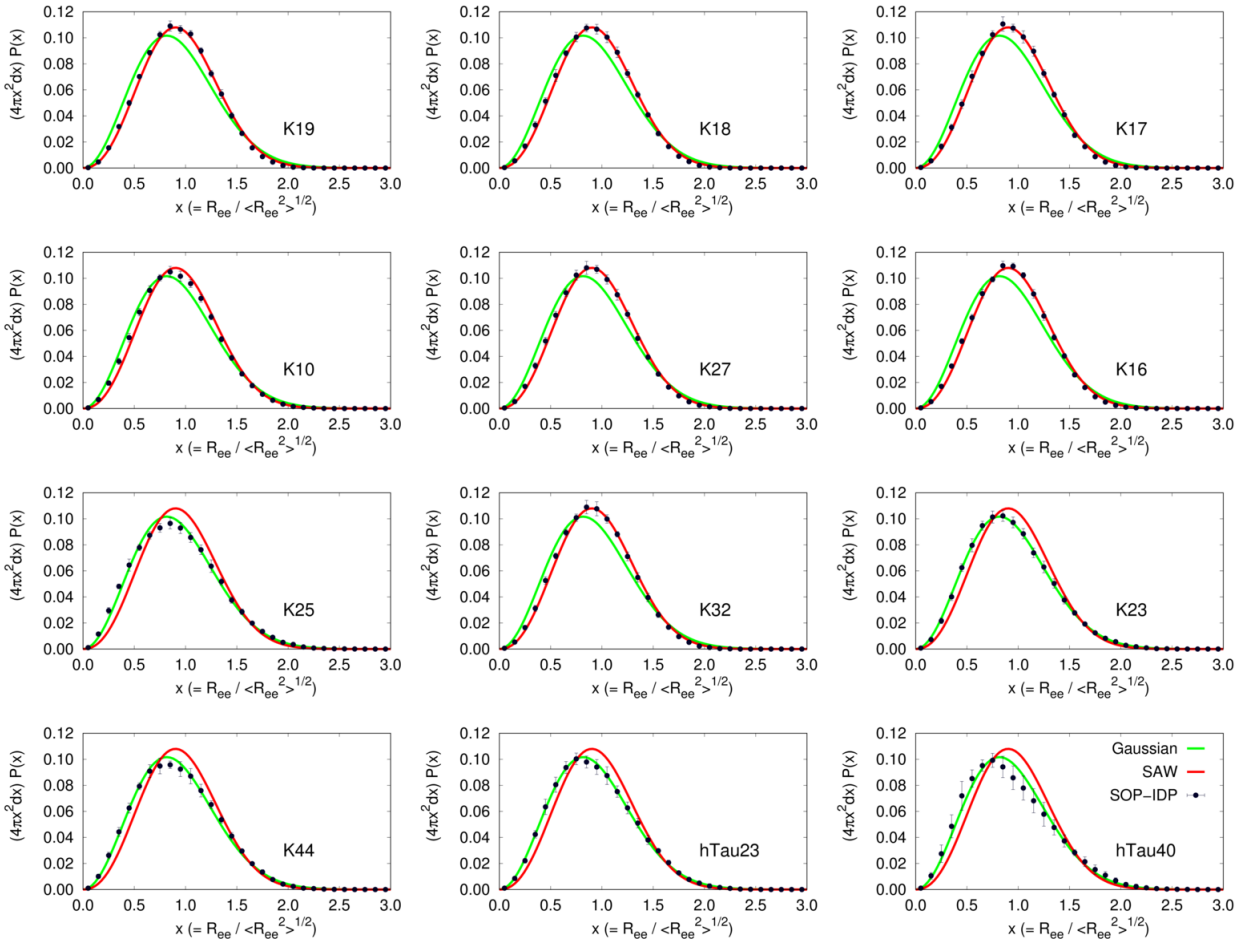


Figure S8: Same as Figure S7, except the results are for for Tau protein constructs. Theoretical distributions for random coil-like behavior and Gaussian chains are shown in red and green, respectively.

Sequence compositional properties

Inspired by sequence variables that naturally arise in theory of polyampholytes (PAs),[8,9] 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, $\frac{(f_+ - f_-)^2}{(f_+ + f_-)}$.[10] 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[11] and PA[8] theories that the phases could be readily altered by external conditions such as changing salt concentration. Clearly, the external conditions are not encoded in f_+ and f_-. Nevertheless, the PA type variable are useful in anticipating the states of IDP. The PA-like properties for the IDP sequences are tabulated in Table S2 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 S3: 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 S10, we show that the WT Tau sequence (hTau40) contains 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 main manuscript for a zoomed-in view, and below for full range), and the K23, K44, and hTau23 constructs (see below).

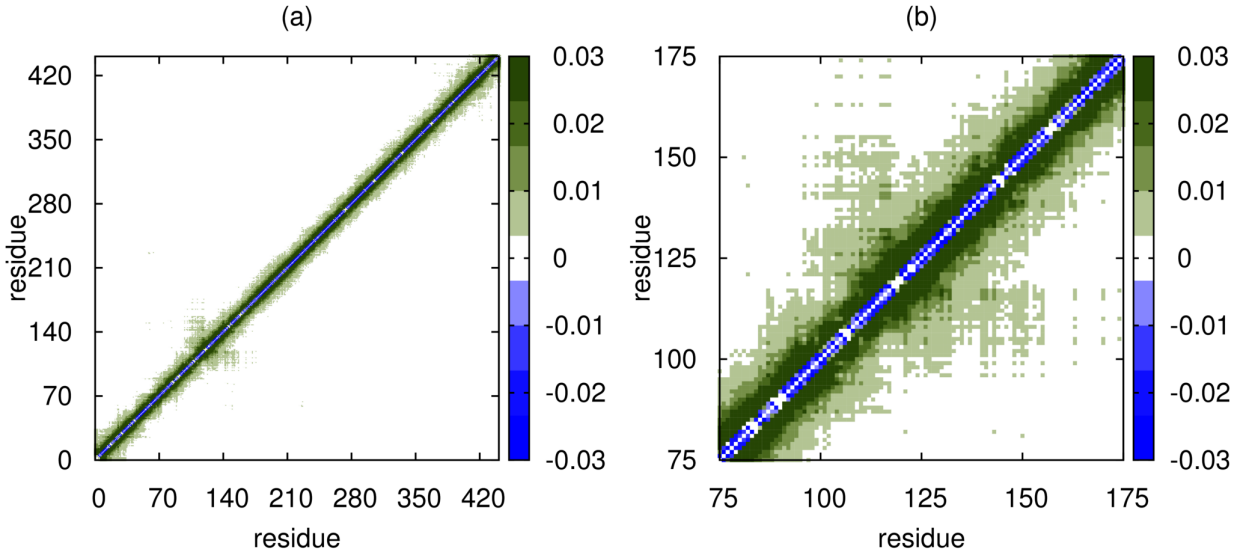


Figure S9: Difference contact map (defined in Eq.1 in the main text) between the one obtained using SOP-IDP simulations and the corresponding contact map for a Flory random coil for (a) the full length Tau protein (hTau40). (b) Zoomed-in view, highlighting the region with enhanced contacts.

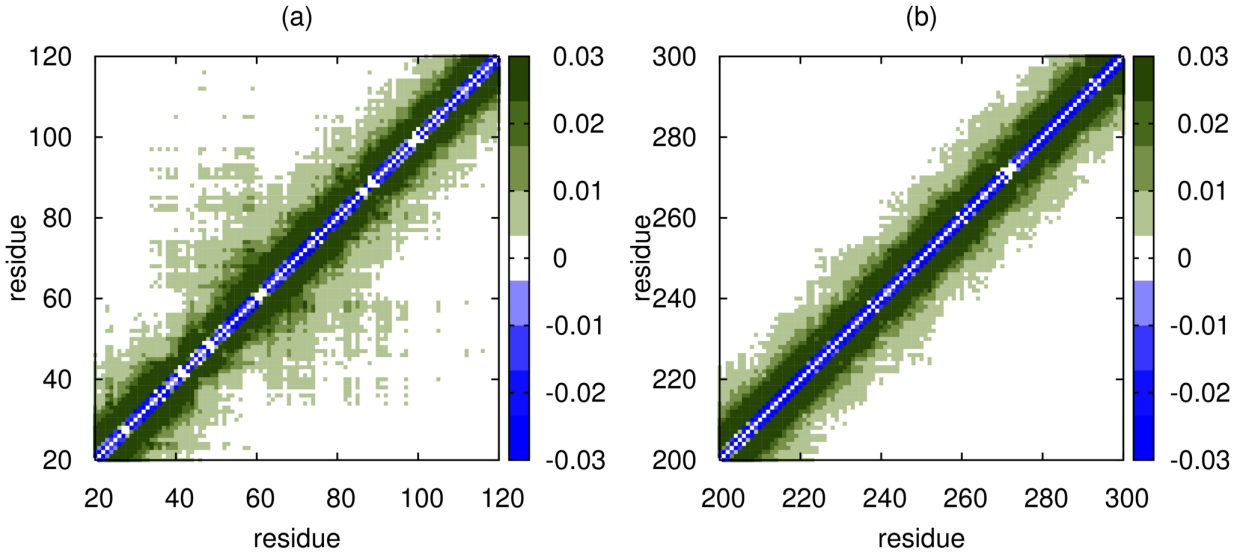


Figure S10: Same as Figure S9 except this if for the hTau23 sequence (a). Zoomed-in view for residues 200-300 of hTau40, shown for comparison (b).

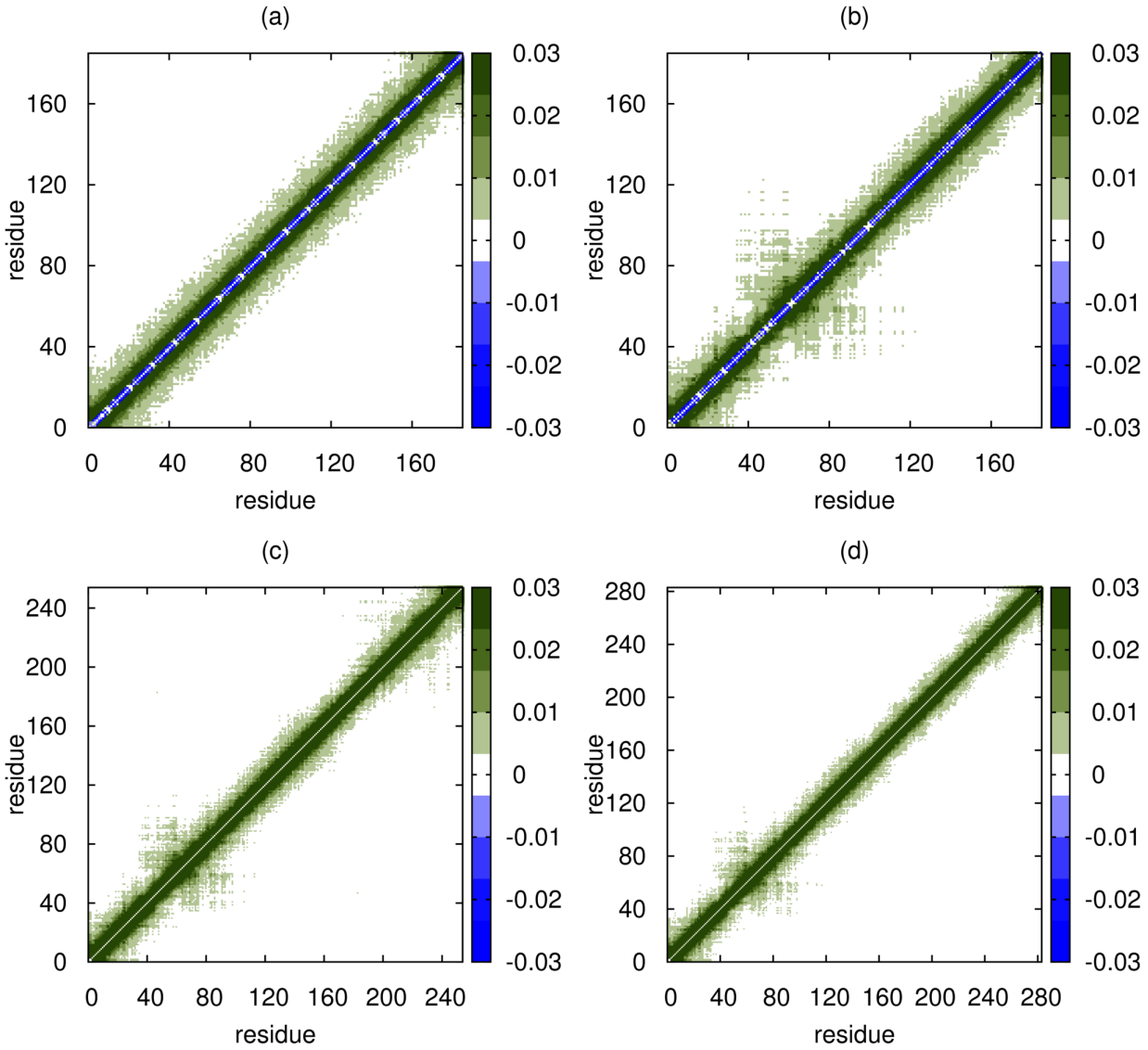


Figure S11: Difference contact maps (Eq. 1 in the main text) for (a) An16, (b) K25, (c) K23 and (d) K44, showing the full range in each case.

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 database DisProt.[12] 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. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 7787–7792.

(2) Reddy, G.; Thirumalai, D. *J. Phys. Chem. B* **2015**, *119*, 11358–11370.

(3) Cragnell, C.; Durand, D.; Cabane, B.; Skepö, M. *Proteins: Structure, Function, and Bioinformatics* **2016**, *84*, 777–791.

(4) Kjaergaard, M.; Nørholm, A.-B.; Hendus-Altenburger, R.; Pedersen, S. F.; Poulsen, F. M.; Kragelund, B. B. *Prot. Sci.* **2010**, *19*, 1555–1564.

(5) Henriques, J.; Skepö, M. *J. Chem. Theory Comput.* **2016**, *12*, 3407–3415.

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

(7) Rubinstein, M.; Colby, R. H. *Polymer physics*; Oxford university press, 2015.

(8) Higgs, P. G.; Joanny, J. *J. Chem. Phys.* **1991**, *94*, 1543–1554.

(9) Ha, B. Y.; Thirumalai, D. *J. Phys. II* **1997**, *7*, 887–902.

(10) Das, R. K.; Ruff, K. M.; Pappu, R. V. *Curr. Opin. Struct. Biol.* **2015**, *32*, 102 – 112.

(11) Ha, B.-Y.; Thirumalai, D. *Phys. Rev. A* **1992**, *46*, R3012–R3015.

(12) Piovesan, D.; Tabaro, F.; Mičetić, I.; Necci, M.; Quaglia, F.; *et. al.*, *Nucleic Acids Res.* **2017**, *45*, D219–D227.